

MapReduce: Simplified Data Processing on Large Clusters (review)

5 December 2019

Reinaldo Astudillo
Delft University of Technology

BigData with Obama



0:00/1:25



MapReduce Schedule

- Background of the MapReduce paradigm
- Concepts in MapReduce
- Examples
- Apache streams
- An introduction to PySpark

Background of the MapReduce paradigm

- Internet explosion at mid's 90.
- Are supercomputer obsolete?
- Cheap clusters.
- New paradigm of calculation needed.
- Big data.

Paper: J. Dean, and S. Ghemawat,

[MapReduce: simplified data processing on large clusters.](https://dl.acm.org/citation.cfm?id=1327492) (https://dl.acm.org/citation.cfm?id=1327492)

Communications of the ACM, Volume 51 Issue 1, 2008

Pages 107-113

Background of the MapReduce paradigm

Functional programming

- Only pure functions (no state variables, no side effects)
- Functions are first class citizen
- Data is immutable

Background of the MapReduce paradigm

MapReduce is based on the functional programming paradigm.

MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.

Concepts in MapReduce

- Mapper: written by the user, takes an input pair and produces a set of intermediate key/value pairs.
- Reducer: accepts an intermediate key and a set of values for that key. Then, it applies the transformation to each value.
- The MapReduce library groups together all intermediate values associated the same key (shuffle and sort phase).

Concepts in MapReduce

Example word count

```
map(string key, string value)
  // key: document name
  // value: document content
  for each word w in value:
    EmitIntermediateKey((w, 1))

reduce(string key, Iterator values)
  // key: a word
  // values: a list of counts
  result = 0
  for each v in values:
    result += v
  Emit((key, result))
```


Concepts in MapReduce

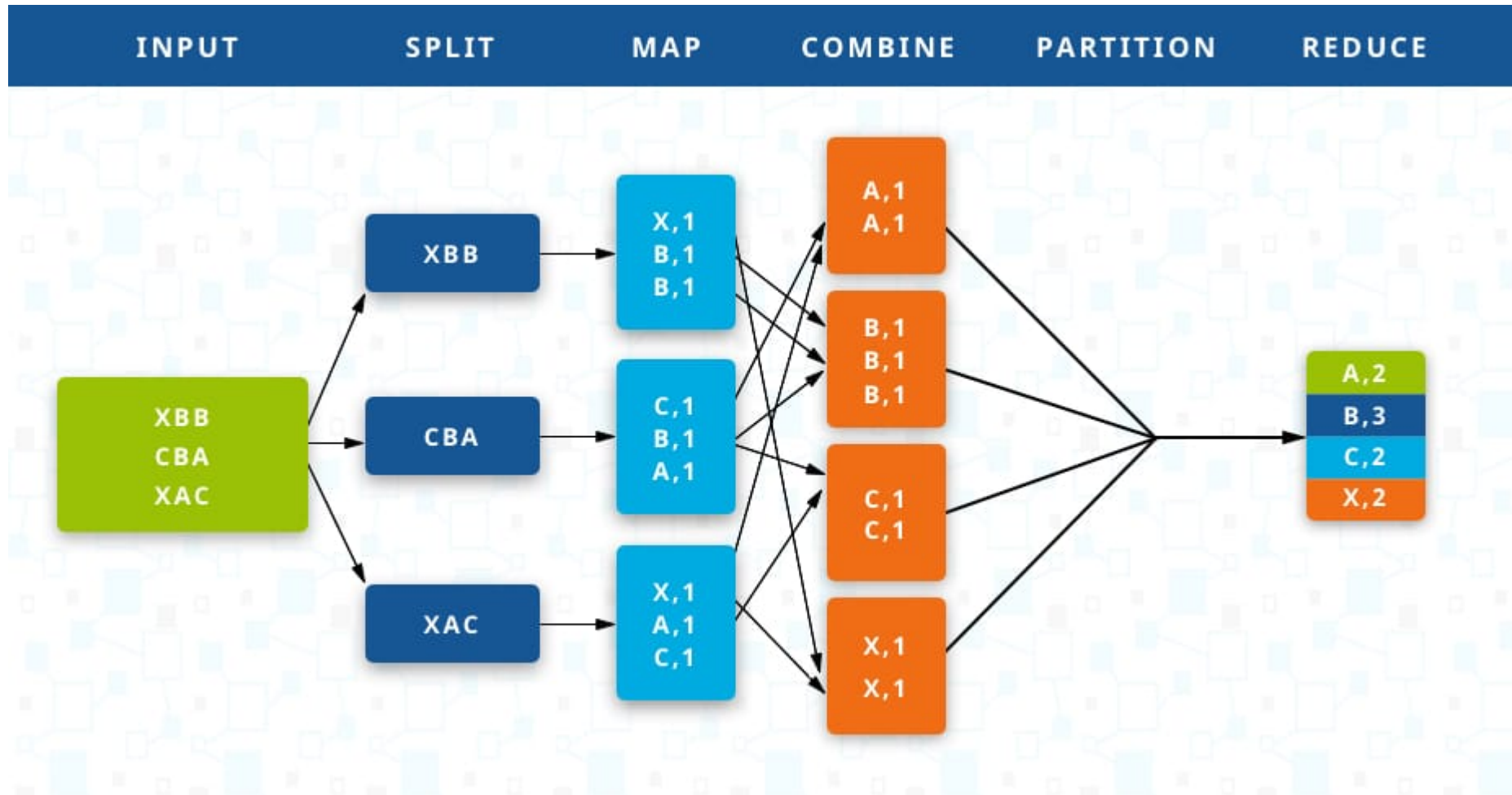


image from <https://www.talend.com/resources/what-is-mapreduce/>

Examples

- Word count
- Distributed sort
- Summarization analytics
- Join operations in distributed data bases (distributed grep, filters)
- A Google type question: Given an array of integers, return a boolean that indicates that this array contains two numbers such that they add up to 100. (Follow up)
What about is this array does not fit in memory?

Thank you

5 December 2019

Tags: Programming, Clusters, BigData, Functional programming, PySpark (#ZgotmplZ)

Reinaldo Astudillo

Delft University of Technology

<http://reinaldoastudillo.nl> (<http://reinaldoastudillo.nl>)

r.a.astudillo@tudelft.nl (<mailto:r.a.astudillo@tudelft.nl>)