

Introduction

Hadoop, HDFS, MapReduce

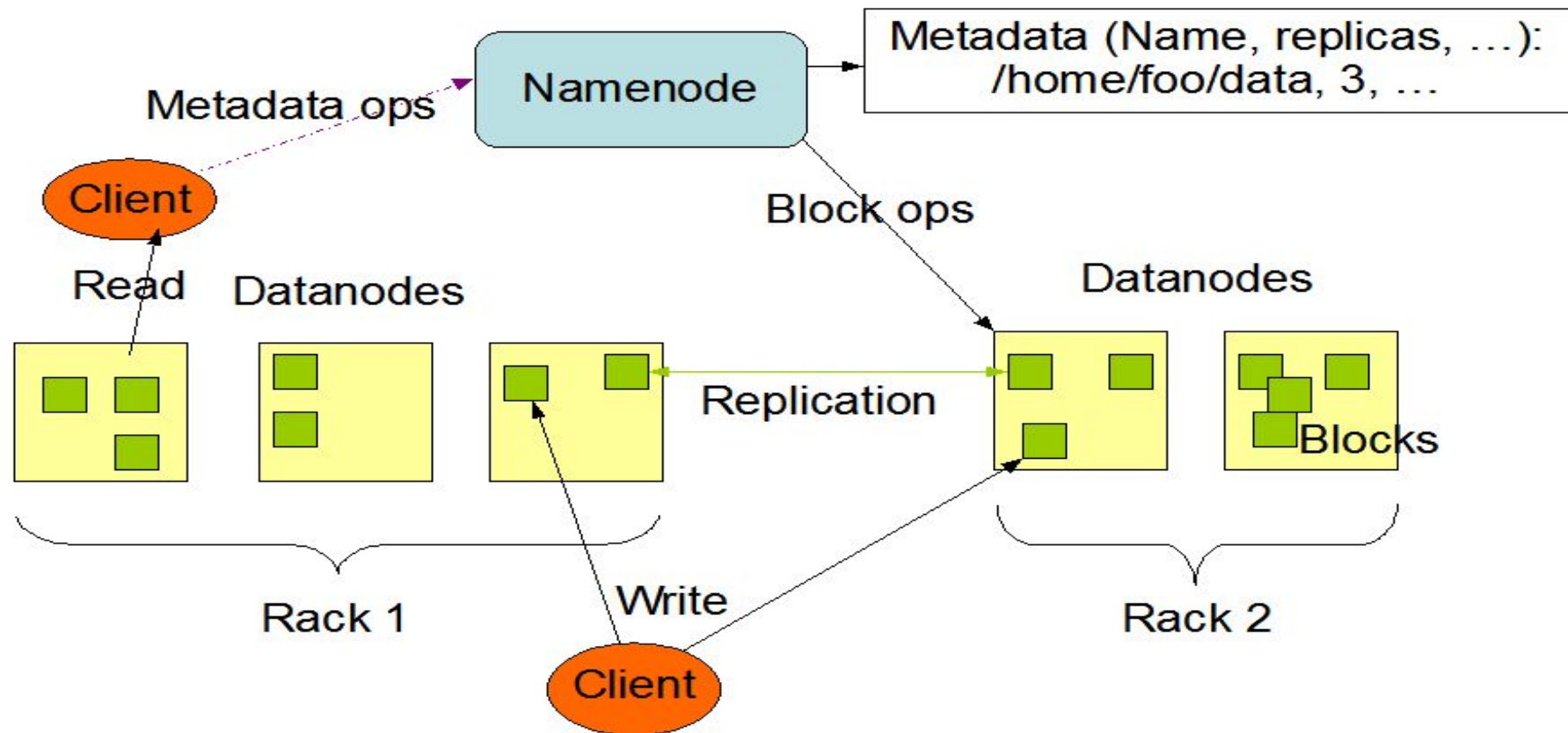
Hadoop

- Open source framework
- Used for storage and large scale processing of data-sets on clusters of commodity hardware
- Mainly consists of the following two modules:
 - HDFS (Distributed Storage)
 - Map Reduce (Analysis/Processing)

HDFS

- Distributed File System based on Google File System (GFS)
- Well suited for commodity hardware
- Built around the idea “Write Once, Read Multiple Times”
- Reliability through replication

HDFS Architecture

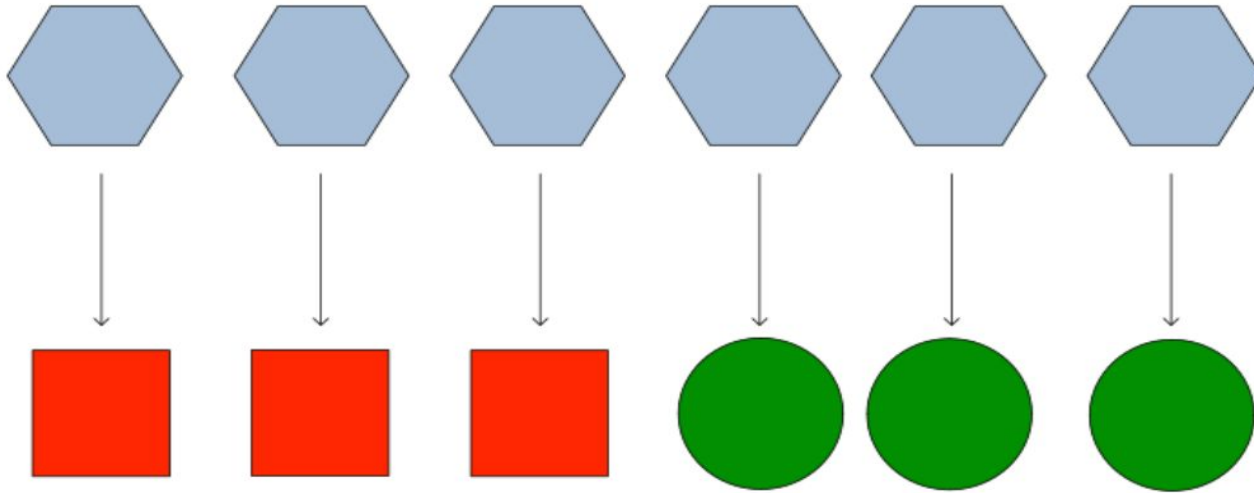


Map-Reduce

- Massively Parallel programming paradigm
- Fault tolerant
- Designed to run on clusters of commodity hardware
- Provides high level of abstraction to developers

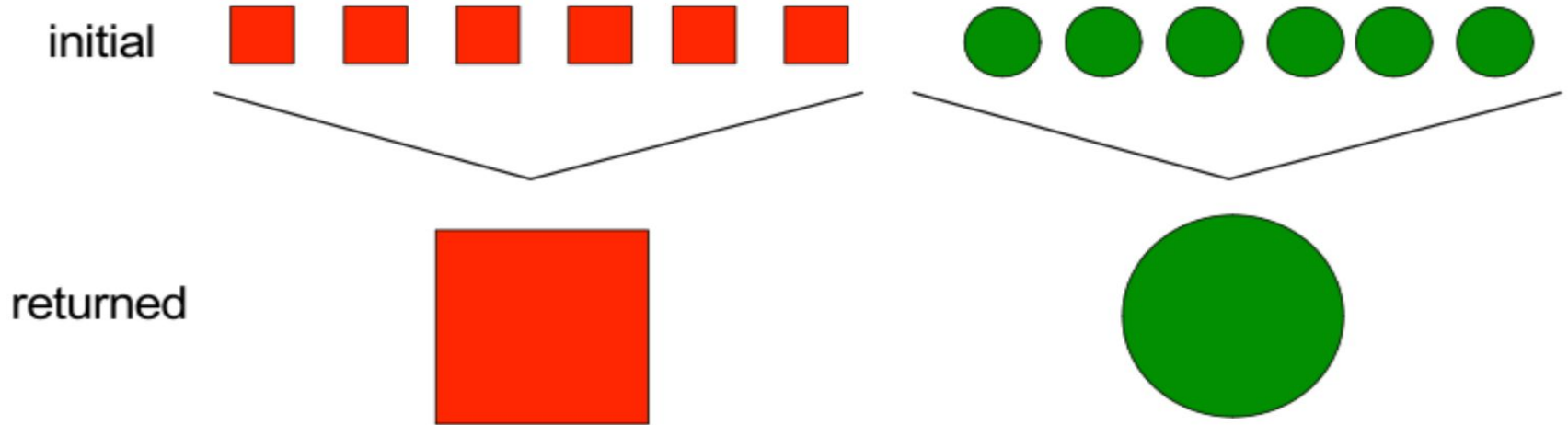
Map

$\text{map}(K1, V1) \rightarrow (K2, V2)$



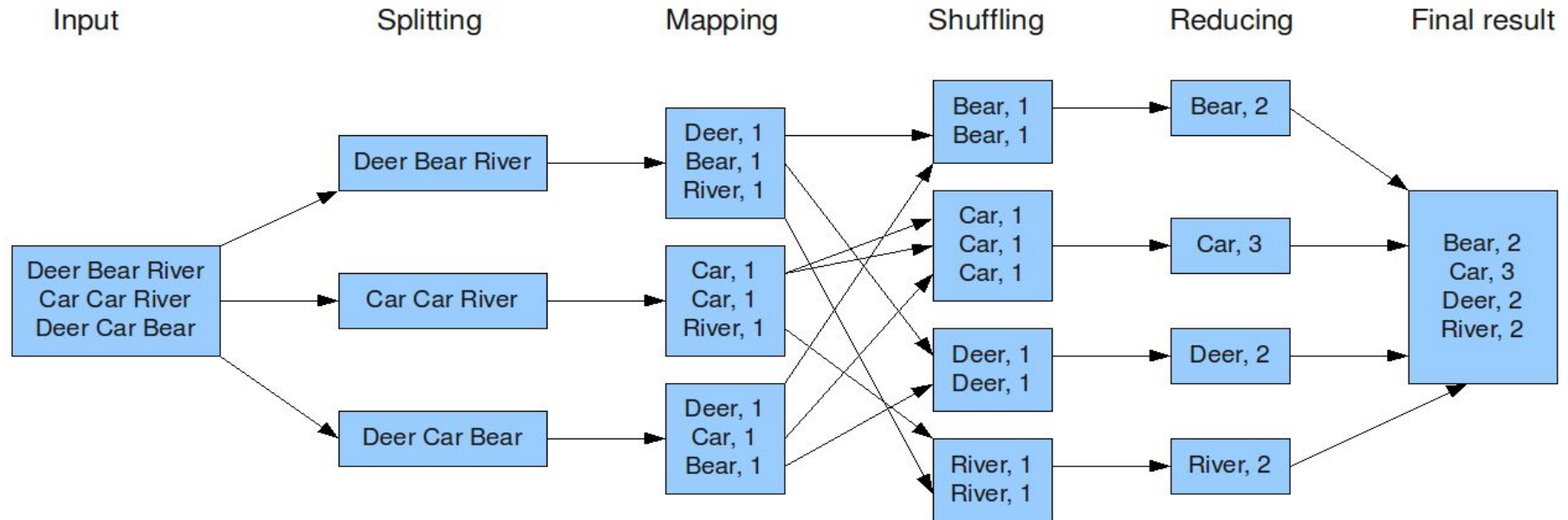
Reduce

`reduce(K2, list(V2) -> list(K3, V3)`

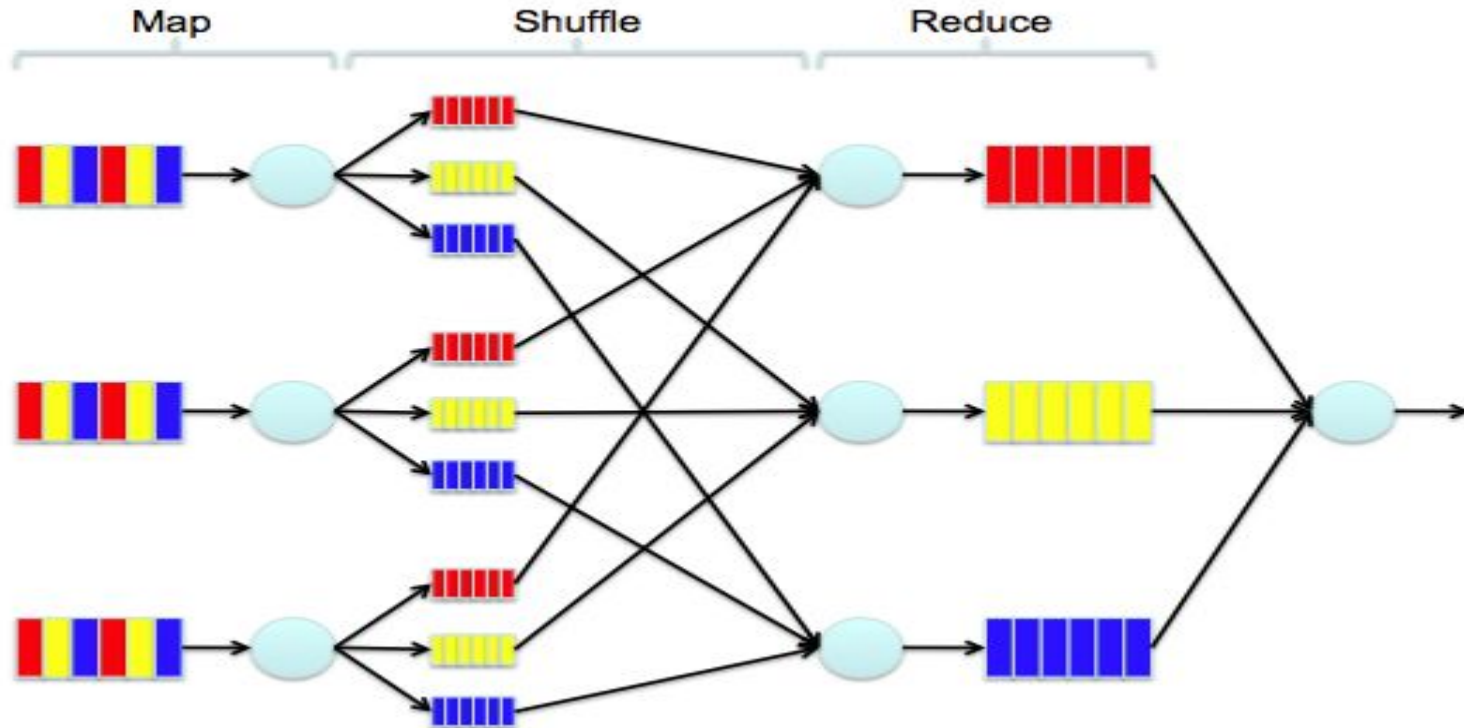


The Famous WordCount

The overall MapReduce word count process



Generalized Flow



A close-up photograph of a person's hand, wearing a dark blue sleeve, adjusting a white slider on a professional audio mixing console. The hand is positioned in the lower-left quadrant of the frame. The mixing console has various knobs and sliders, with some red and blue components visible on the right side. The background is heavily blurred, showing out-of-focus lights in shades of green, yellow, and red, creating a bokeh effect. A semi-transparent dark grey horizontal band spans the middle of the image, containing the text "Let's Code It Now! :)". A thin vertical green line is positioned to the left of the text.

Let's Code It Now! :)