IIT KHARAGPUR

NPTEL ONLINE
CERTIFICATION COURSES

NPTEL

Cloud Computing : *MapReduce - Tutorial*

Prof. Soumya K Ghosh

Department of Computer Science and Engineering

IIT KHARAGPUR

Introduction

- MapReduce: programming model developed at Google
- Objective:
  - Implement large scale search
  - Text processing on massively scalable web data stored using BigTable and GFS distributed file system
- Designed for processing and generating large volumes of data via massively parallel computations, utilizing tens of thousands of processors at a time
- Fault tolerant: ensure progress of computation even if processors and networks fail
- Example:
  - Hadoop: open source implementation of MapReduce (developed at Yahoo!)
  - Available on pre-packaged AMIs on Amazon EC2 cloud platform

MapReduce Model

- Parallel programming abstraction
- Used by many different parallel applications which carry out large-scale computation involving thousands of processors
- Leverages a common underlying fault-tolerant implementation
- Two phases of MapReduce:
  - Map operation
  - Reduce operation
- A configurable number of M ‘mapper’ processors and R ‘reducer’ processors are assigned to work on the problem
- The computation is coordinated by a single master process

MapReduce Model Contd...

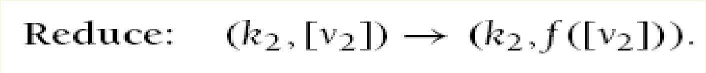
- Map phase:
  - Each mapper reads approximately *1/M* of the input from the global file system, using locations given by the master
  - Map operation consists of transforming one set of key-value pairs to another:

$$\textbf{Map:}\quad (k_1, v_1) \rightarrow [(k_2, v_2)].$$

  - Each mapper writes computation results in one file per reducer
  - Files are sorted by a key and stored to the local file system
  - The master keeps track of the location of these files

MapReduce Model Contd…

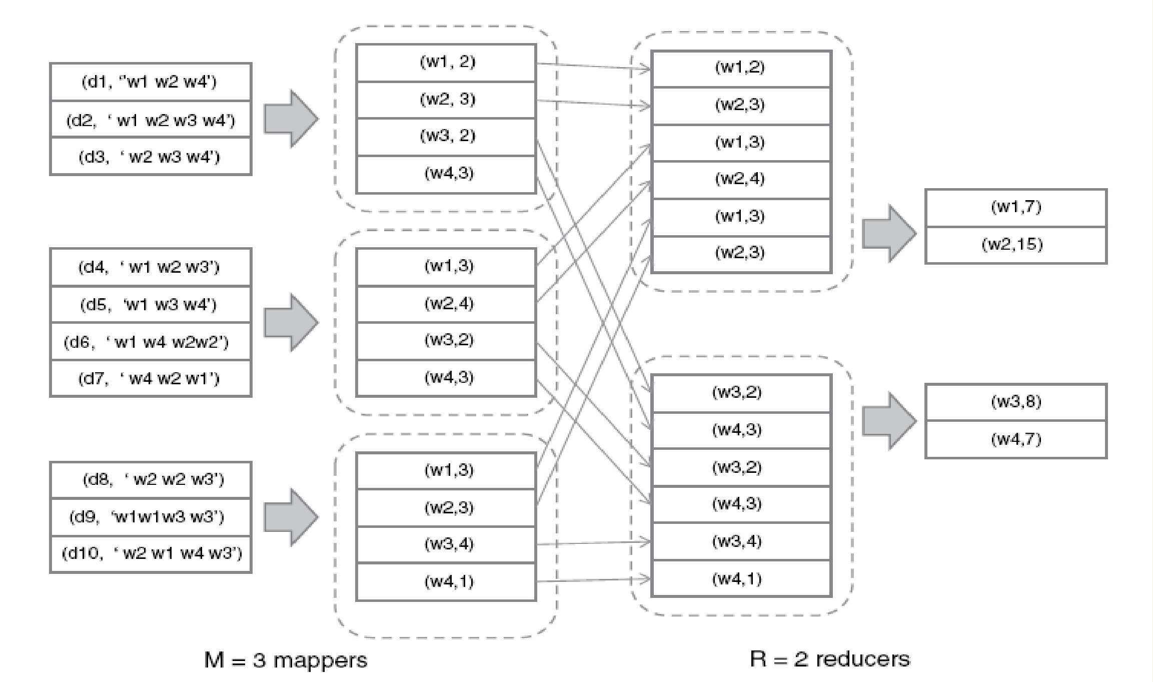
- **Reduce phase:**
  - The master informs the reducers where the partial computations have been stored on local files of respective mappers
  - Reducers make remote procedure call requests to the mappers to fetch the files
  - Each reducer groups the results of the map step using the same key and performs a function *f* on the list of values that correspond to these key value:



$$\text{Reduce:} \quad (k_2, [v_2]) \rightarrow (k_2, f([v_2])).$$

  - Final results are written back to the GFS file system

MapReduce: Example



(d1, "w1 w2 w4')
(d2, ' w1 w2 w3 w4')
(d3, ' w2 w3 w4')

(d4, ' w1 w2 w3')
(d5, 'w1 w3 w4')
(d6, ' w1 w4 w2w2')
(d7, ' w4 w2 w1')

(d8, ' w2 w2 w3')
(d9, 'w1w1w3 w3')
(d10, ' w2 w1 w4 w3')

(w1, 2)
(w2, 3)
(w3, 2)
(w4,3)

(w1,3)
(w2,4)
(w3,2)
(w4,3)

(w1,3)
(w2,3)
(w3,4)
(w4,1)

(w1,2)
(w2,3)
(w1,3)
(w2,4)
(w1,3)
(w2,3)

(w3,2)
(w4,3)
(w3,2)
(w4,3)
(w3,4)
(w4,1)

(w1,7)
(w2,15)

(w3,8)
(w4,7)

M = 3 mappers

R = 2 reducers

- 3 mappers; 2 reducers
- Map function:

$$(d_k, [w_1 \ldots w_n]) \to [(w_i, c_i)].$$

- Reduce function:

$$(w_i, [c_i]) \to \left(w_i, \sum_i c_i \right)$$

Problem-1

In a MapReduce framework consider the HDFS block size is 64 MB. We have 3 files of size 64K, 65Mb and 127Mb. How many blocks will be created by Hadoop framework?

Problem-2

Write the pseudo-codes (for map and reduce functions) for calculating the average of a set of integers in MapReduce.

Suppose A = (10, 20, 30, 40, 50) is a set of integers. Show the map and reduce outputs.

Problem-3

Compute total and average salary of organization XYZ and group by based on gender (male or female) using MapReduce. The input is as follows

Name, Gender, Salary

John, M, 10,000

Martha, F, 15,000

Problem-4

Write the *Map* and *Reduce* functions (pseudo-codes) for the following ***Word Length Categorization*** problem under *MapReduce* model.

Word Length Categorization: Given a text paragraph (containing only words), categorize each word into following categories. Output the frequency of occurrence of words in each category.

Categories:

tiny: 1-2 letters; **small:** 3-5 letters; **medium:** 6-9 letters; **big:** 10 or more letters

Thank You!