MapReduce:
Simplified Data Processing on Large Clusters
(Woohyung Han, Youngdeok Seo)
Outline

What is MapReduce?
MapReduce Architecture
MapReduce Example
Summary
Motivation
Motivation

- A Single machine can't serve all data
  - need distributed system
- Parallel programming
  - Threading is hard!
    - how to communicate between nodes?
    - how to scale out?
    - how to handle machine failures?

Solution: MapReduce!
What is MapReduce?

- Pioneered by Google
- Parallel **programming model** meant for large clusters
  - Parallelization
  - Fault Tolerance
  - Data Distribution
  - Load Balancing
- User only implements Map() and Reduce()
What is MapReduce used for?

- At google
  - Index construction for Google Search
  - Article clustering for Google News
  - Statistical machine translation
- At Yahoo!
  - Yahoo! Search
  - Spam detection for Yahoo! Mail
- At Facebook
  - Data Mining
  - Spam Detection
Outline

What is MapReduce?
MapReduce Architecture
MapReduce Example
Summary
Challenges

- **Node failure**
  - Mean time between failures for 1 node = 1 years
  - Mean time between failures for 10k nodes = 53 min
  - **Solution**: Fault-tolerance

- **Network traffic**
  - **Solution**: Push computation to the data

- **Programming distributed system is hard**
  - **Solution**: Make Data-parallel programming model that well defined.
MapReduce workflow

Input Data

Split 0
Split 1
Split 2

Worker
read

Worker
local write

Worker
remote read, sort

Map
extract something you care about from each record

Reduce
aggregate, summarize, filter, or transform

Output Data

Write

Output File 0

Output File 1
Map()

- Inputs a key/value pair
  - Key is a reference to the input value
  - Value is the data set on which to operate
- Evaluation
  - Function defined by user
  - Applies to every value in value input
    - Might need to parse input
def map(key, value):
    list = []
    for x in value:
        if test:
            list.append((key, x))
    return list
Reduce()

- Accepts the Mapper output
- Aggregates value on the key
  - Merge all intermediate values associated with the same key
def reduce(key, listOfValues):
    result = 0
    for x in listOfValues:
        result += x
    return (key, result)
Outline

What is MapReduce?
MapReduce Architecture
MapReduce Example
Summary
MapReduce Example

• Word Count
  – Count how many words include
    • To be, or not to be, that is the question.
      – to: 2
      – be: 2
      – or: 1
      – ...

This is a book. That book is on the desk.
I like that book.
Outline

What is MapReduce?
MapReduce Architecture
MapReduce Example

Summary
Summary

• We introduced MapReduce
  − programming model for processing large scale data

• MapReduce provides:
  − a general-purpose model to simplify large-scale computation
  − Allows user to focus on the problem without worrying about details