Chapter 1

Introducing Haoop
Preface

• We’re surrounded by data.
• The exponential growth of data first presented challenges to cutting-edge businesses.
  – Terabytes, petabytes
• Existing tools were becoming inadequate to process such large data sets
  – MapReduce
Preface (Cont’d)

• Leading universities, such as Standford and CMU, have already started introducing Hadoop into their computer science curriculum.
What is Hadoop?

• Hadoop is an open source framework for writing and running distributed applications that process large amounts of data

• Key distinctions of Hadoop are
  – Accessible
  – Robust
  – Scalable
  – Simple
Understanding distributed system and Hadoop

• Distributed system (scale-out) and huge monolithic servers (scale-up)
  – Refer to 2th paragraph

• Hadoop differs from schemes such as SETI@home
Comparing SQL databases and Hadoop

• Many of Hadoop’s initial applications deal with unstructured data such as text.
• Scale-out instead of scale-up
• Key/value pairs instead of relational tables
• MapReduce instead of SQL
  – Scripts and codes VS query statements
Understanding MapReduce

• MapReduce is also a data processing model
  – Easy scaling of data processing over multiple computing nodes
  – Data processing primitives
    • Mappers and Reducers
Scaling a simple program manually

• Pseudo code in page 9
  – Tokenize (method)
  – wordCount (array)

• However to process large documents?
  – Refer to 2th para. in page 10.

• Two phases to solve the problem
  – Refer to page 11 (last para.)
Scaling the same program in MapReduce

• MapReduce programs are executed in two main phases
  – Mapping and reducing
Counting words with Hadoop—running your first program

- Linux is the official development and production platform for Hadoop
- Running Hadoop requires Java
- Hadoop script
  - Refer to page 15