# Discussion Session
# Week 6

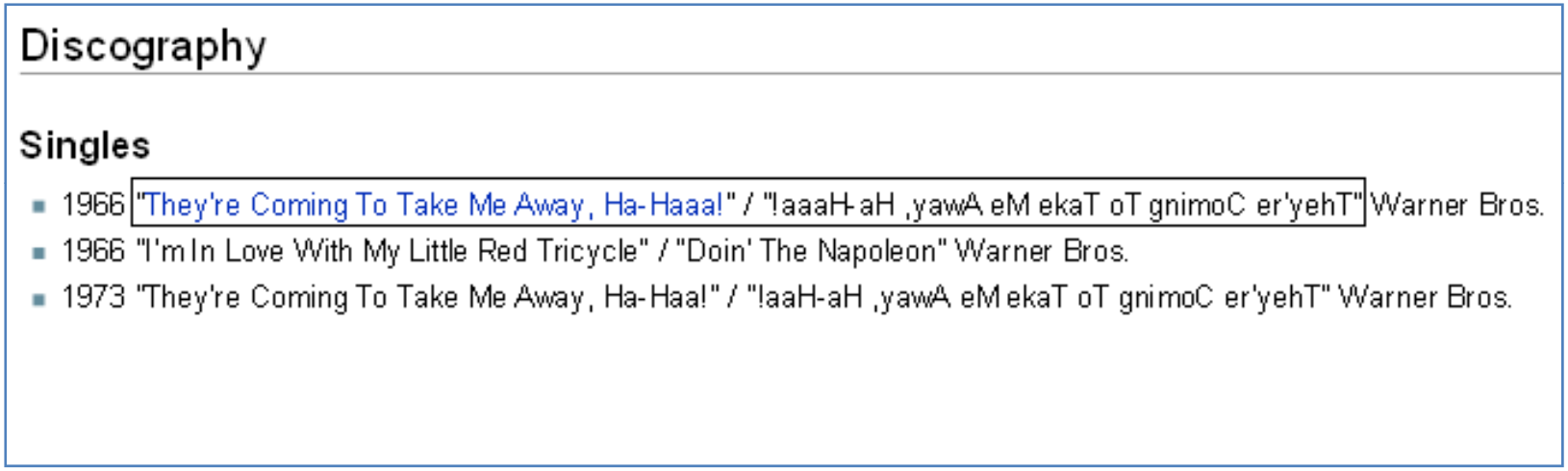
INF 141: Information Retrieval
Winter 2008

Yasser Ganjisaffar

yganjisa@ics.uci.edu

# Assignment 03

- Longest known (to me) Palindrome in Wikipedia:



Discography

**Singles**

- 1966 "They're Coming To Take Me Away, Ha-Haaa!" / "!aaaH-aH ,yawA eM ekaT oT gnimoC er'yehT" Warner Bros.
- 1966 "I'm In Love With My Little Red Tricycle" / "Doin' The Napoleon" Warner Bros.
- 1973 "They're Coming To Take Me Away, Ha-Haa!" / "!aaH-aH ,yawA eM ekaT oT gnimoC er'yehT" Warner Bros.

- Source: http://en.wikipedia.org/wiki/Napoleon_XIV

- Longest reported Palindrome:
  - s. tropeR trebloC ehT si sihT." ("This is The Colbert Report" s

# Assignment 03

- Longest known (to me) Rhopalic in Wikipedia (17 words):
- Source: [http://en.wikipedia.org/wiki/ Lexicant](http://en.wikipedia.org/wiki/ Lexicant)

V
IV
TIV
ITIV
SITIV
NSITIV
NSITIVI
NSITIVIT
ENSITIVIT
ENSITIVITI
ENSITIVITIE
SENSITIVITIE
OSENSITIVITIE
OSENSITIVITIES
POSENSITIVITIES
YPOSENSITIVITIES
**HYPOSENSITIVITIES**

- Longest reported Rhopalic:
  - of ABN Amro Hoare Govett, Tourism Training Australia
  - and hair salon, called "Dalohan," becoming customers themselves

# Assignment 03 – Your Results

| Pages | Links | Duration | Text | Content | Rhopalic | Lipogram | Palindrome | Obama | Link/Page | HTML/Text |
|---|---|---|---|---|---|---|---|---|---|---|
| 912,500 | 137,338,042 | 2.5d | 9.5 | 41.3 | 7 | OK-vandalism | OK-vandalism | Not Found | 150.5074433 | 4.347368421 |
| 233,921 | 79,486,017 | 2.5d | 3.3 | 13.4 | 7 | Mid | Short - Not vandalism | 4946 | 339.7985516 | 4.060606061 |
| 100,000 | 23,164,593 | 6.4h | 1.6 | 6.4 | 6 + bug | OK-Source? | | | 231.64593 | 4 |
| 235,712 | 60,394,129 | ? | **4.4** | **7.8** | 5 | Too Short-Not vanda | Very Short - Not vand | **1** | 256.2200015 | 1.772727273 |
| | | 7h | | | 8 | OK | OK - All a -vandalism | | | |
| 321,000 | 59,358,861 | 1d | | 4.2 | 6 | OK | OK-vandalism | 3340 | 184.9185701 | |
| 17,000 | 4,926,699 | 4h | 0.4 | | 5 | Mid | Mid | 3486 | 289.8058235 | |
| 800,000 | 123,857,058 | 2.5d | 8.6 | 37.5 | 8 | OK-vandalism | OK-vandalism | 47 | 154.8213225 | 4.360465116 |
| 445,950 | 74,909,376 | 3d | 5.4 | 22.5 | 7 | OK | OK (Peel's foe not a se | 2120 | 167.9770737 | 4.166666667 |
| 360,000 | 61,636,580 | 2d | 4.6 | 18.7 | 5 | Mid+ | OK-non sense | 40772 | 171.2127222 | 4.065217391 |
| 137,500 | 27,755,110 | 2d (6h?) | 2.1 | 8.3 | 6 | OK | --- | 2 (seeded) | 201.8553455 | 3.952380952 |
| 1,012,148 | 147,904,209 | 2d | 10.7 | 46.6 | 8 | OK-vandalism | OK - All a -vandalism | 2353 | 146.1290335 | 4.355140187 |
| 845,200 | **1.27X10^10???** | 12h (24h?) | | | 7 | OK | OK | 3340 | | |
| 37,273 | 8,353,413 | 12h | 0.6 | 2.4 | 7 + bug | OK | OK (Peel's foe not a se | **2** | 224.1143187 | 4 |
| 737 | 218,646 | 11h | | | --- | Bad | --- | | 296.6702849 | |
| | | | | | 4 | Short - Source? | Lost | | | |
| 259,550 | 46,478,416 | 3d | 3.3 | 13.6 | 8 | OK | OK (s. tropeR trebloC | 17213 | 179.0730726 | 4.121212121 |
| **202,789** | **1,770,923** | 6d | 2.2 | 10 | 8 | OK | OK | 25480 | 8.732835607 | 4.545454545 |

# What is Hadoop

Flexible infrastructure for large scale computation and data processing on a network of commodity hardware.

# Current State of Hadoop Project

- Top level Apache Foundation project
- In production at Yahoo!, Facebook, IBM, Amazon, Fox, NY Times, …
- Very active and strong development team

- Last year, Yahoo! Launched a Hadoop application that runs on a more than 10,000 core Linux Cluster and produces data that is now used in every Yahoo! Web search query.

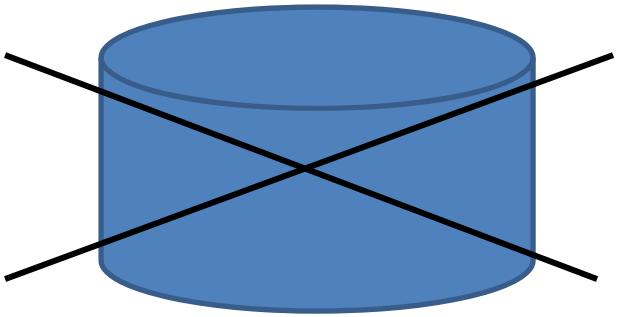
# What is Hadoop?

The Linux of Distributed Processing.

# Very Large Storage Requirements

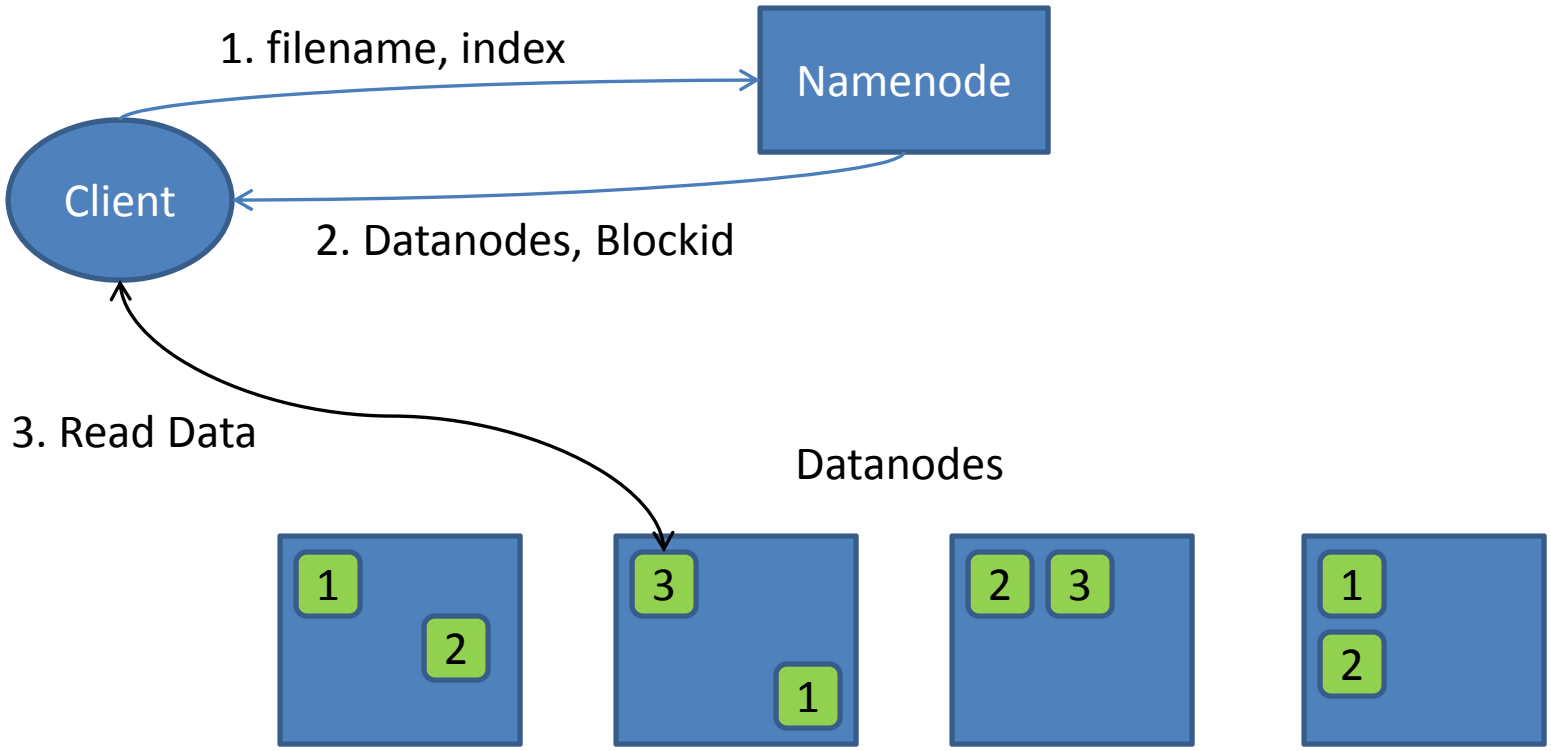
10,000,000 GB

No single storage can handle this amount of data.

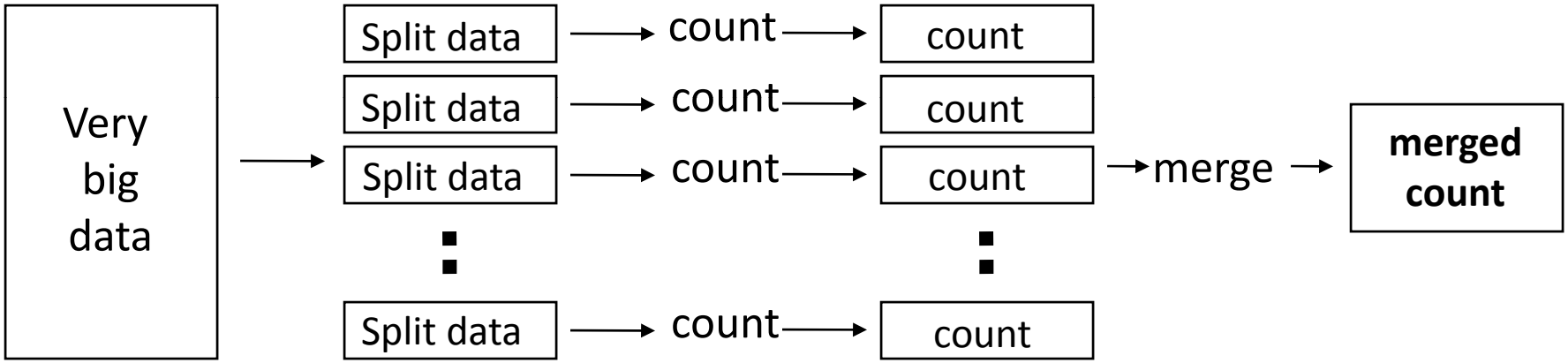


You need a large set of nodes each storing part of the data.

# HDFS


1. filename, index
Namenode
Client
2. Datanodes, Blockid
3. Read Data
Datanodes
1
2
3
1
2
3
1
2

# Distributed Word Count

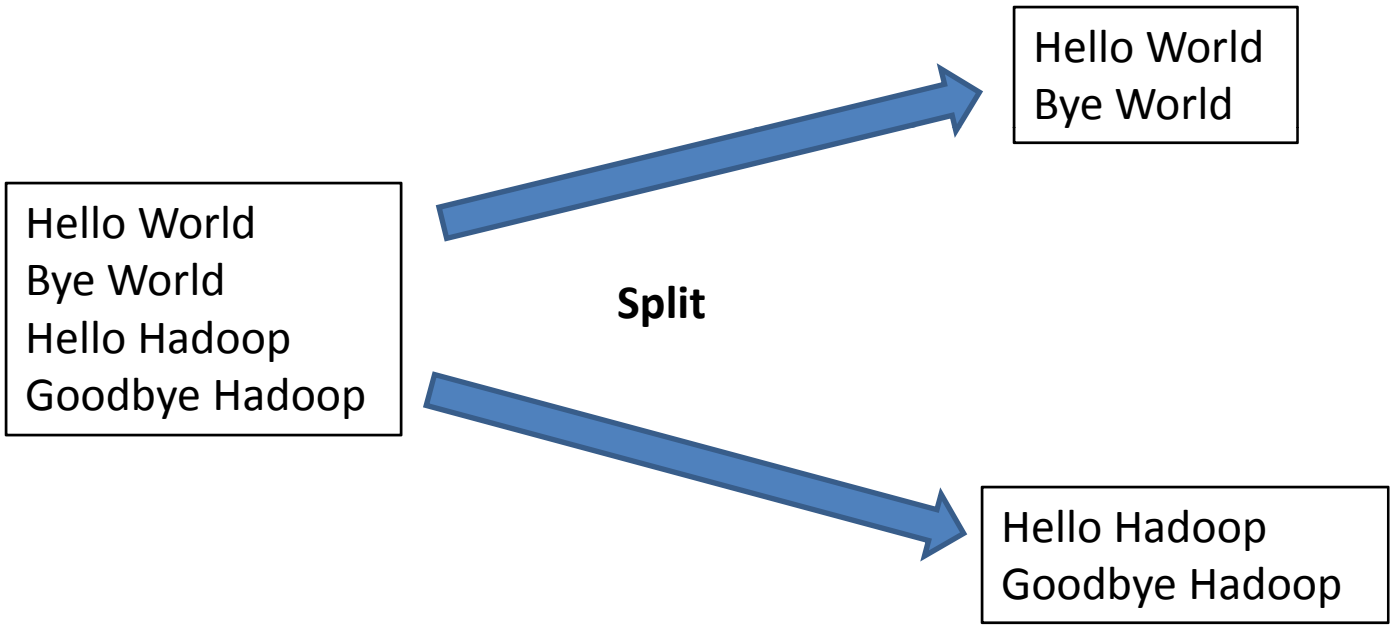


# Programming Model: Map/Reduce

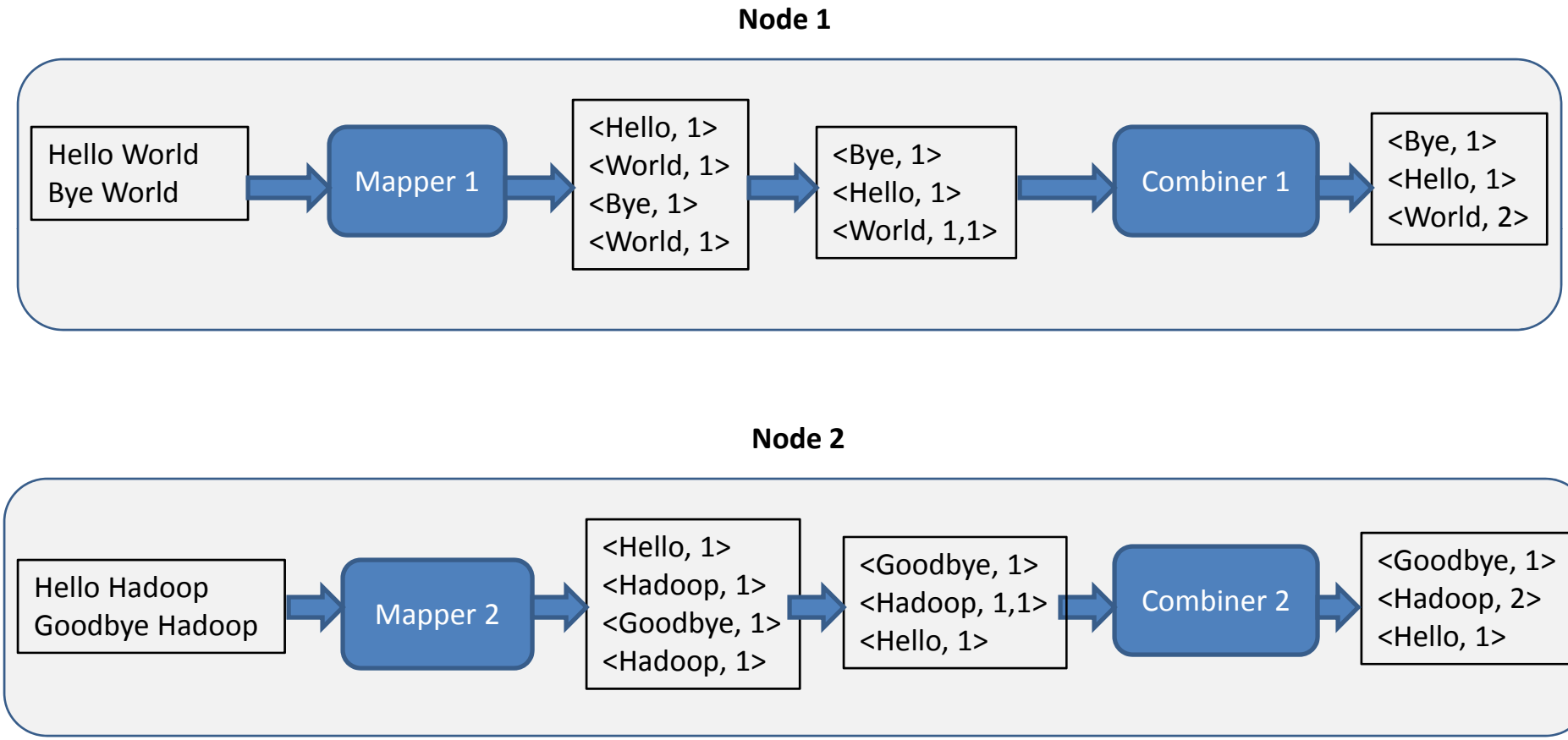
- Very simple programming model:
  - Map(anything)->key, value
  - *Sort, partition on key*
  - Reduce(key,value)->key, value

- No parallel processing / message passing semantics

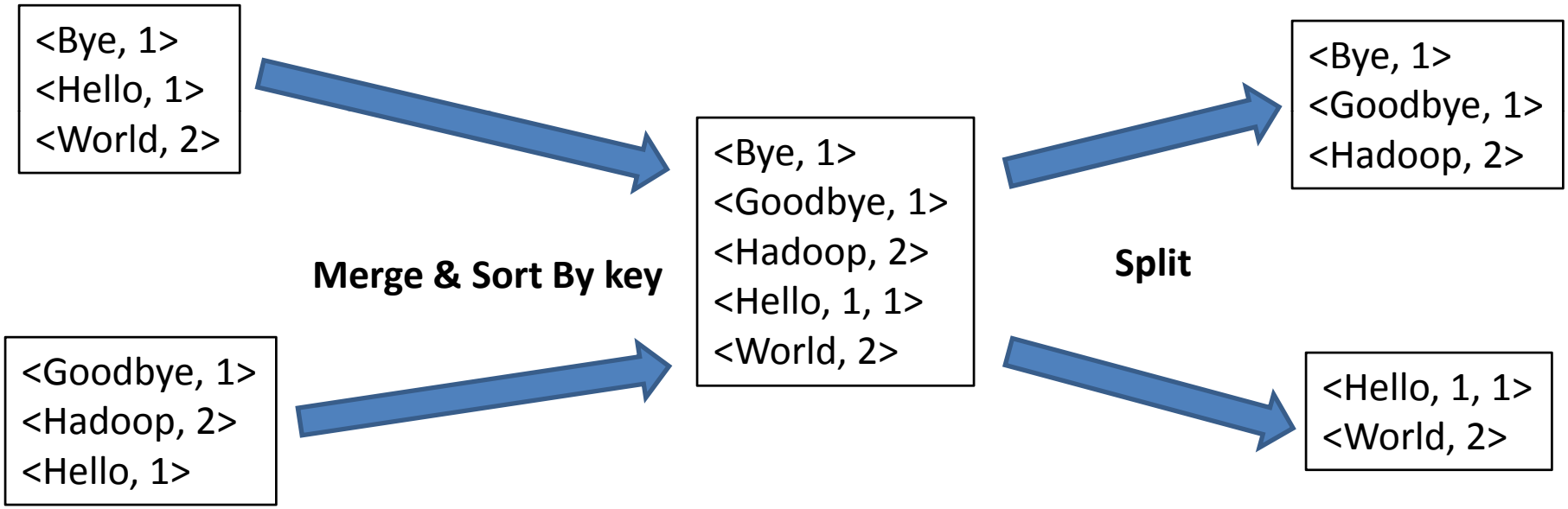
# Counting Words by Map/Reduce



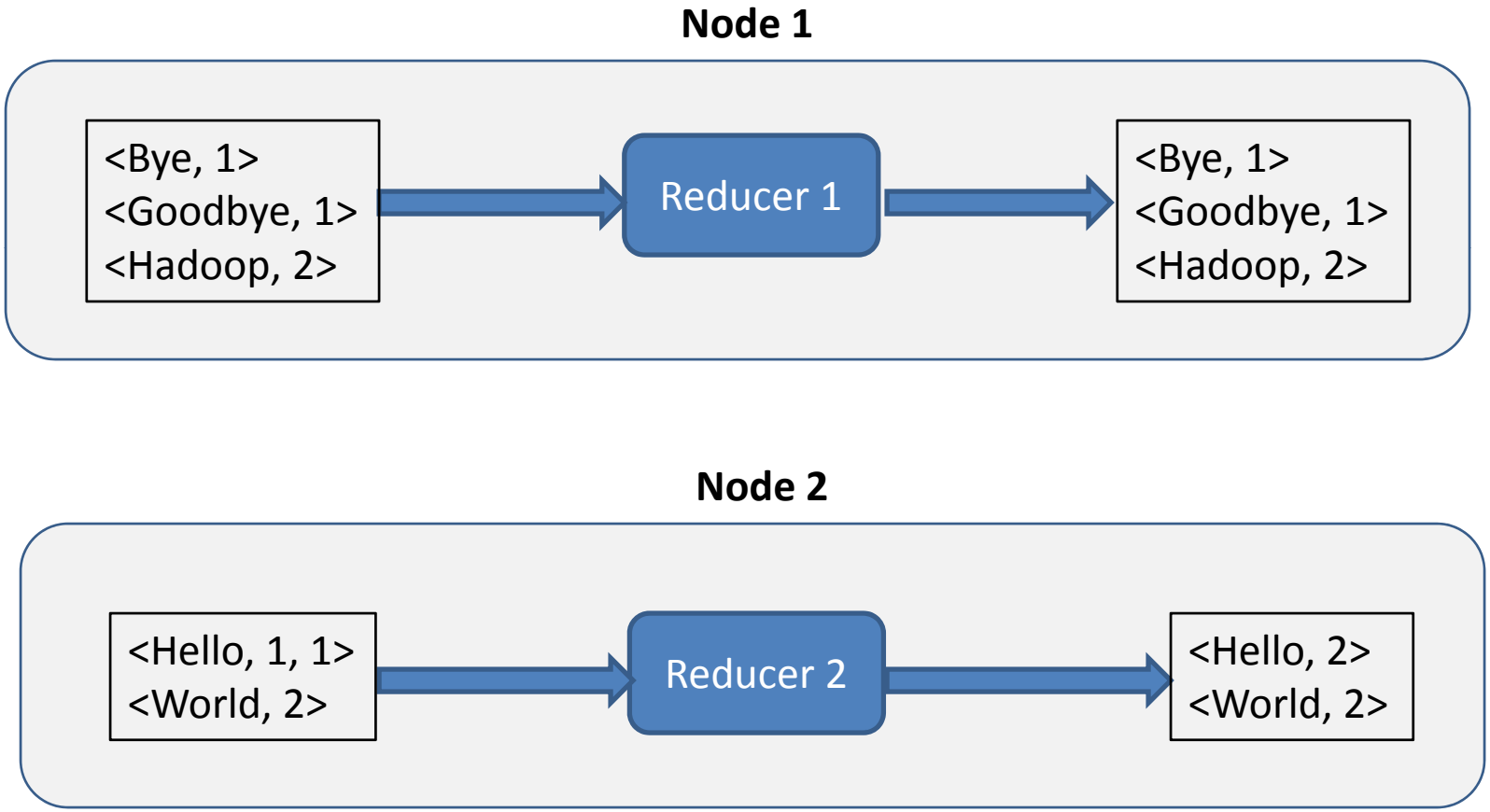
# Counting Words by Map/Reduce



# Counting Words by Map/Reduce



# Counting Words by Map/Reduce



# Installation Requirements

- Linux
- Java 1.6.x
- ssh

- Download from:
  - http://hadoop.apache.org/core/releases.html

# Links

- MapReduce: Simplified Data Processing on Large Clusters, http://labs.google.com/papers/mapreduce.html
- Google File System: http://labs.google.com/papers/gfs.html