# Comparative Study of HPCC and HADOOP based on PUMA Benchmarks

CSC 591-007 : Data Intensive Computing
Instructor: Dr. Vincent W. Freeh
Group No.2

Presentation by:
Haiyu Yao
Vishal Mishra
Nakul Shukla

# Introduction

- Develop and run benchmark scripts for HPCC and Hadoop
  - Hadoop, a popular framework for processing large data.
  - HPCC, a computing platform to process and deliver big data solutions.
  - PUMA is a benchmark suite for Hadoop.

# Motivation

- Lack of comparison results based on different types of problems.
- Published available comparisons.
  - Based on TeraSort between HPCC and Hadoop
- Last benchmark test on HPCC was performed in 2011 on Terasort.

# Benchmark Scripts Used

- Word Count
- Inverted Index
- Adjacency - List
- Self - Join
- K-means Clustering
- Histogram - Movies

# Environment

- Virtual Machine Cluster
  - 1 master node
  - 8 slave nodes
- Main Memory: 2 GB per node (usually 1 GB free)
- CPU Cores: 2 per node (Intel(R) Xeon(R) E5645 @ 2.40GHz)
- Storage: 30 GB per node
- Operating System: Ubuntu Linux 14.04 Base

# Environment (Cont.)

- Hadoop
  - dfs.blocksize: 128 MB (64 MB for Adjacency-List)
  - dfs.replication: 1
  - mapreduce.map.java.opts: 512 MB
  - mapreduce.reduce.java.opts: 768 MB
  - mapreduce.task.io.sort.mb: 256MB
- HPCC
  - 2 thor slave processes per node
  - no roxie
  - replication factor: 2 (default configuration)

# Methods

- Same Input/Output
- Same cluster environment
- Consecutively run same benchmark task for Hadoop and HPCC
- Execute each benchmark task at least twice
- Test different size of data

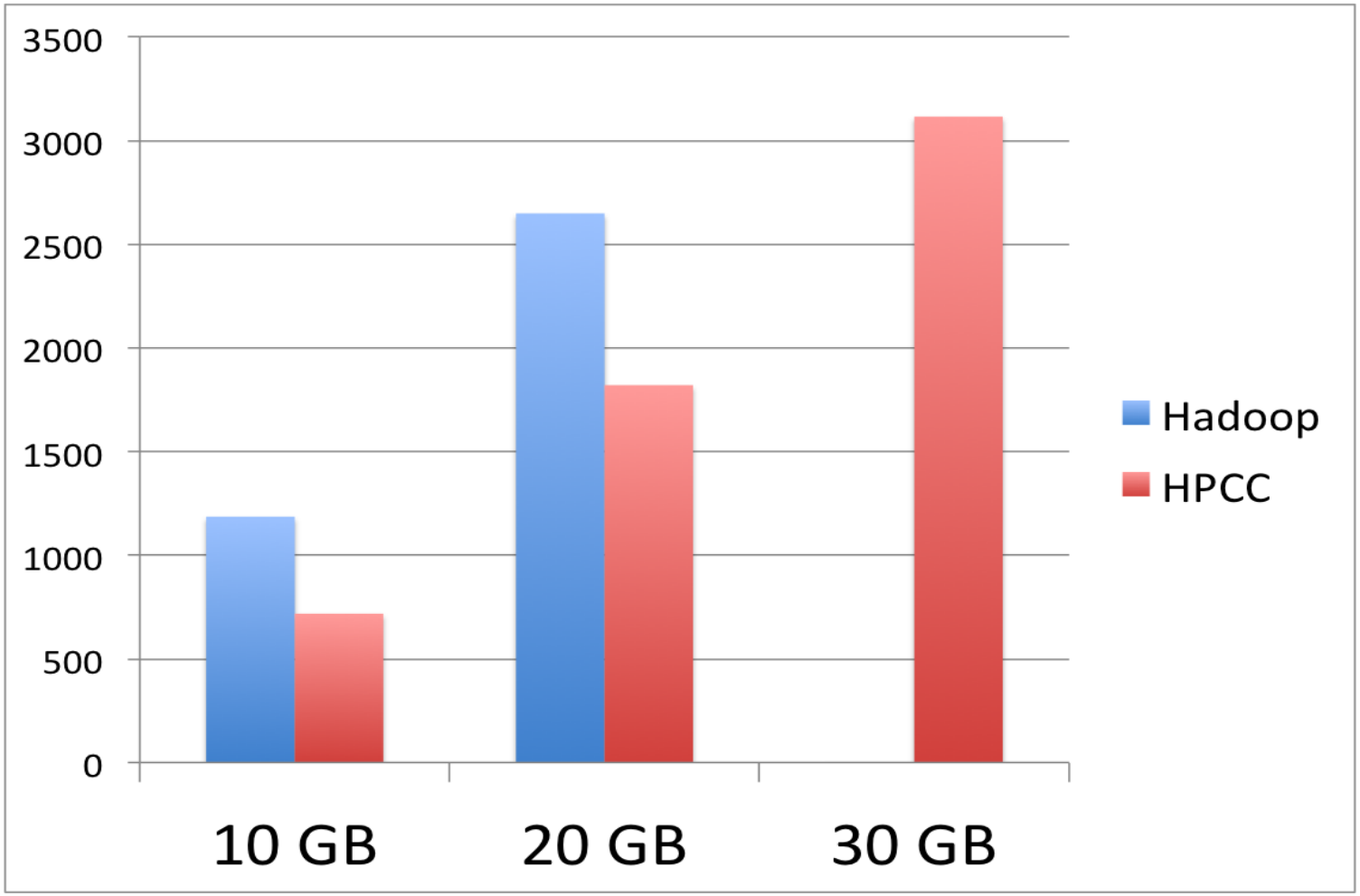
# Results

Unit: Minute

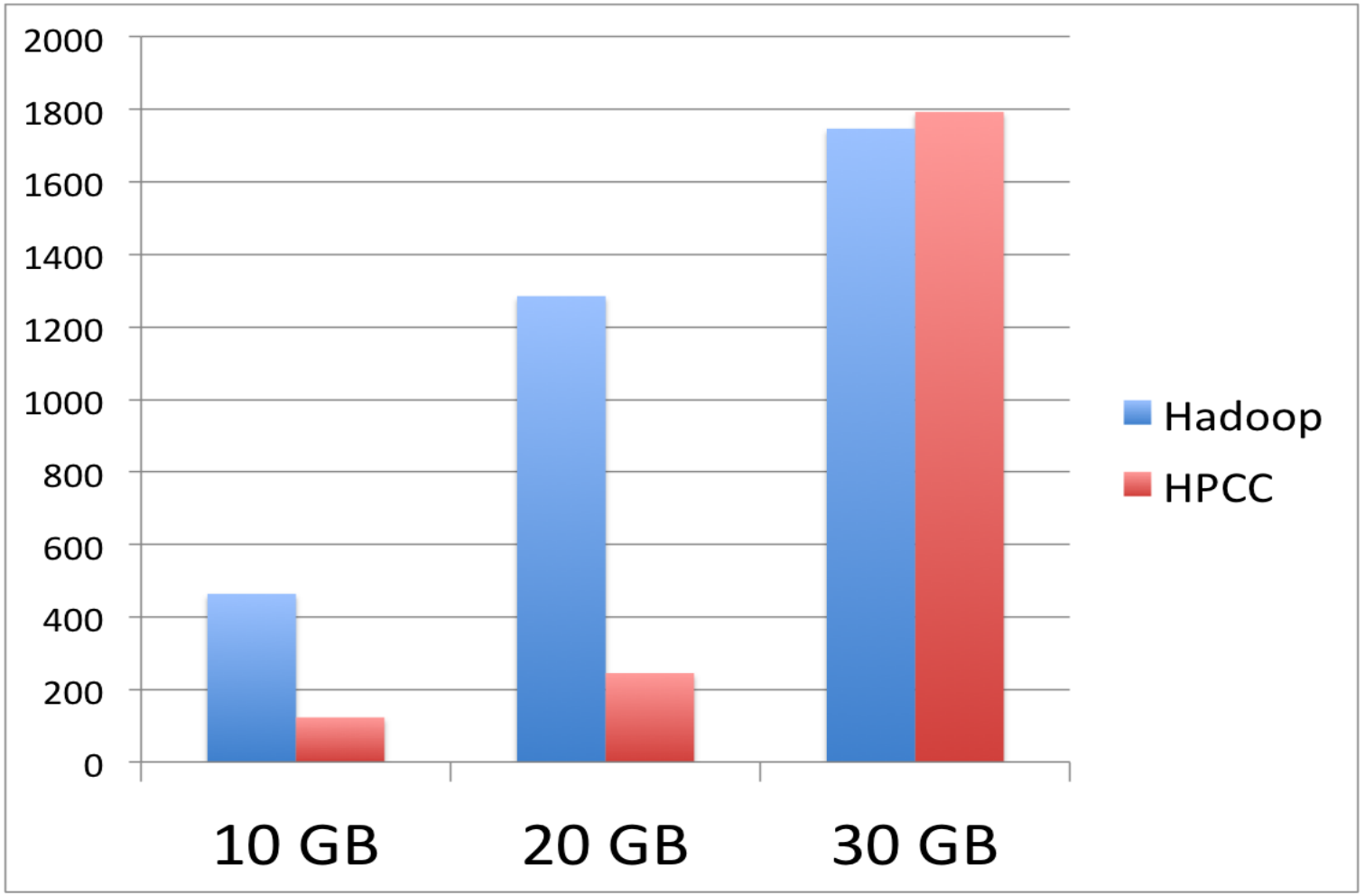
| | 10 GB | | 20 GB | | 30 GB | |
|---|---|---|---|---|---|---|
| | Hadoop | HPCC | Hadoop | HPCC | Hadoop | HPCC |
| Word Count | 7:42 | 2:03 | 21:25 | 4:05 | 29:07 | 29:53 |
| Inverted Index | 10:01 | 6:19 | 21:49 | 10:52 | 27:42 | 16:22 |
| Adjacency-List | 19:46 | 11:58 | 44:09 | 30:20 | N/A | 51:56 |
| Self-Join | 4:58 | 5:46 | 11:20 | 13:52 | 16:01 | 21:28 |

| | 1 GB | | 10 GB | | 30 GB | |
|---|---|---|---|---|---|---|
| | Hadoop | HPCC | Hadoop | HPCC | Hadoop | HPCC |
| K-Means | 00:36 | 2:14 | 3:16 | 13:45 | 8:17 | 34:19 |
| Histogram | 00:36 | 19:56 | 2:59 | 3:02:57 | 7:27 | 9:08:51(Est.) |

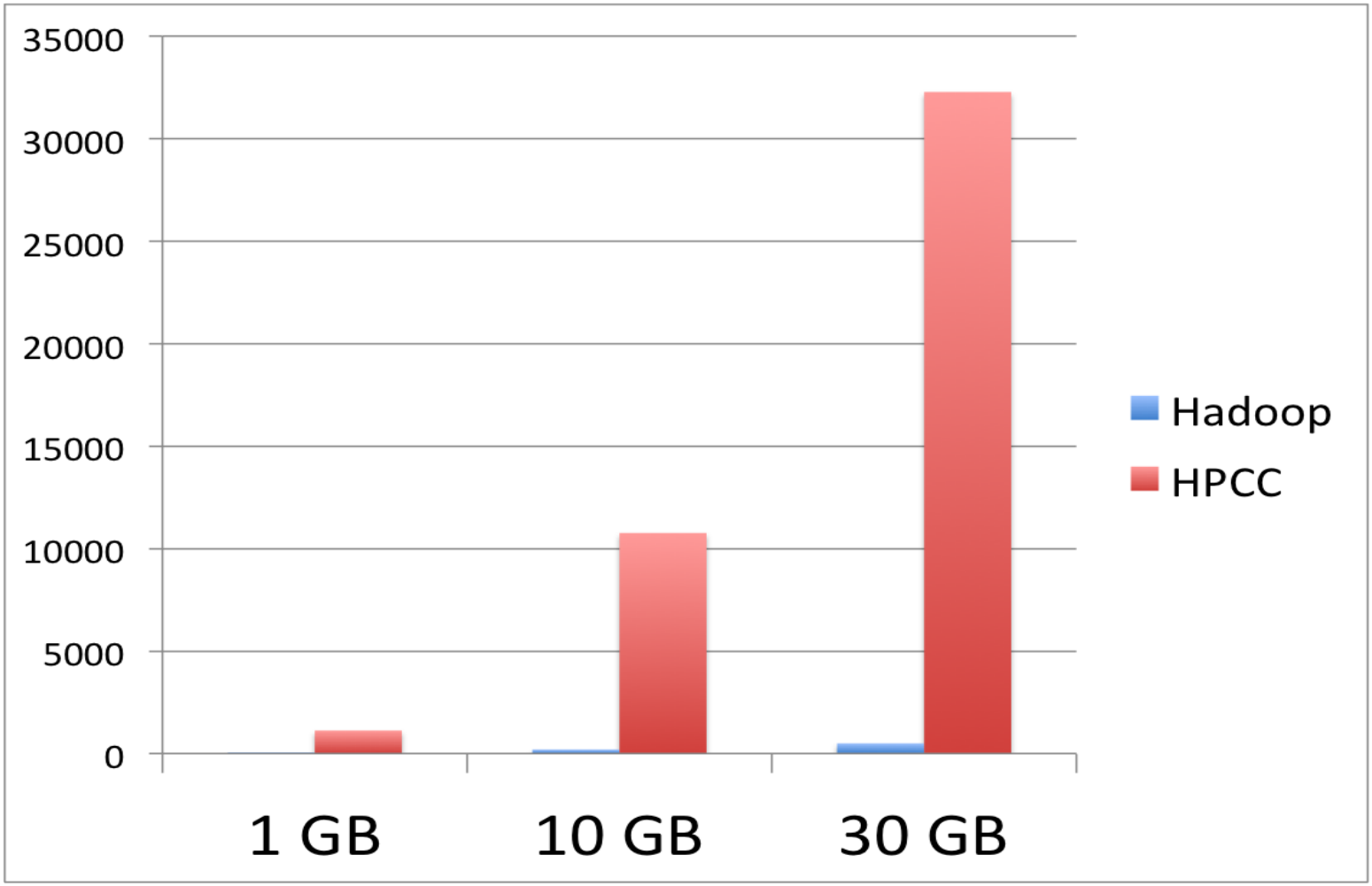
# Interesting Facts (Adjacency-List)



# Interesting Facts (Word Count)



# Interesting Facts (Histogram)



# Lessons Learned (HPCC)

- Pros
  - HPCC is quite fast when less data pre-processing required
  - HPCC is easy to deploy
  - Works well with non- embarrassingly independent tasks
  - Easier to work with different distributions of data
- Cons
  - HPCC is not ideal for data pre-processing
  - Learning curve of ECL is steep
  - ECL is difficult to debug

# Lessons Learned (Hadoop)

- Pros
  - Faster for data pre-processing
  - Flexible in dealing with data in different formats
  - Good community support and documentation
- Cons
  - Configuration is complicated
  - Sensitive to memory shortage
  - Requires embarrassingly independent tasks

# Q & A