Lecture 1: MapReduce: Counting at Scale
Modeling Social Data, Spring 2017
Columbia University

Samuel Meshoyrer

February 10, 2017

1 What is Hadoop?

- An abstraction for parallel data analysis
- consists of many subprojects

  - we will mostly focus on Map/Reduce
- deals with distributed data
- born out of open source web indexing, crawling, and searching software
- Map/Reduce revealed by Google in 2004

  - added to Hadoop, which is adopted by Yahoo! in 2006

2 Why do we need Map/Reduce?

- There is still a lot of latency when dealing with a lot of data
- Read speed of commodity hard disk is about 1 TB/4hrs
- Using Hadoop, 1PB can be sorted in 16.5 hours! petabytesort

3 Map/Reduce

- break into parts
- process in parallel
- combine results

For example, if we wanted to count the number of occurences of each word in a book:

1. For every word on every page
2. Map to (word, count) e.g. (”cat”, 1)
3. Shuffle to collect all records with the same key (word)

   hash(val) = hashVal mod number of reducers
4. Reduce results by adding count values for each word

4 Introduction

This is where your text goes. If you're new to LaTeX, check out Overleaf[1], an online LaTeX environment where you can edit and render your documents. They also have a very useful getting started guide.

Figure 1 is an example of how to include an image.

Figure 1: This is how to include a figure. As long as you use pdflatex most file types (e.g., jpg, png, pdf) should work.

And here's some math:

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi} \tag{1}$$

You can also make numbered lists:

1. Thing 1
2. Thing 2
3. etc.

Or bulleted lists

- Thing 1
- Thing 2
- etc.

It shouldn't get much more complicated than that.

[1] http://overleaf.com