Special Topics: Big Data Lecture 1 About STBD

Gregory S. DeLozier, Ph.D.

gdelozie@kent.edu

Mission

- Learn about the basics of big data methods
  - Specifically map-reduce in detail

Mission

- Learn about the basics of big data methods
  - Specifically map-reduce in detail
- Learn about setting up a big data system
  - Tools and languages
  - Performance considerations

Mission

- Learn about the basics of big data methods
  - Specifically map-reduce in detail
- Learn about setting up a big data system
  - Tools and languages
  - Performance considerations
- Learn some data science skills
  - Working with big data sets
  - Statistical analysis & machine learning

Mission

- Learn about the basics of big data methods
  - Specifically map-reduce in detail
- Learn about setting up a big data system
  - Tools and languages
  - Performance considerations
- Learn some data science skills
  - Working with big data sets
  - Statistical analysis & machine learning
- Do a big (>1M) data analysis project
  - Present methods, results, visualizations, etc.

Mission, Secondary

- Learn how to write technical proposals/papers
- Learn how to create presentations
- Examine some research in big data topics
- Learn how to present results on the web
- Impress your friends

- Enjoy yourself!

Not In Scope

- Learning how to program computers
  - Get some Java and Python skills
  - Get some web and small database skills
- General software development skills
- Fear of breaking things

General Approach

- We will be using industry documentation
  - Tutorials
  - Videos
- We will be learning how to use things
- Examples in GitHub, open source data
- Making something with the technology
- Examination, homework, etc.

Grading (Since you asked…)

- Weekly homework. Do this. 20%
- Attendance. Get to class. 10%
- One class project. 30%
- One research paper. 20%
- Final exam. Open notes, etc. 20%

Grading (Since you asked…)

- Weekly homework. Do this. 20%
- Attendance. Get to class. 10%
- One class project. 30%
- One research paper. 20%
- Final exam. Open notes, etc. 20%

- Plagiarism -100%

Policies – Plagiarism

- Plagiarism will not be tolerated.
- Read the Kent State University policy.
- Visit http://www.plagiarism.org

Realities – Plagiarism

- Plagiarism doesn't work.
- Plagiarism makes you seem desperate.
- Plagiarism can get you fired.
- Plagiarism usually makes me disagreeable.

Realities – Plagiarism

- …but not always. Sometimes it's amusing.

"Specifically, it is vital that information be finished, right, and a la mode as for the outer world."

Realities – Plagiarism

- …but not always. Sometimes it's amusing.

"Specifically, it is vital that information be finished, right, and a la mode as for the outer world."

Realities – Plagiarism

- …but not always. Sometimes it's amusing.

"Specifically, it is vital that information be finished, right, and a la mode as for the outer world."

("***a la mode"*** **:** served with ice cream on it)

Pie a la mode

Realities – Plagiarism

- …and then there's this.

"Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size:"

Realities – Plagiarism

- …and then there's this.

"Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size:"

"Bigtable is a conveyed stockpiling framework for overseeing organized information that is intended to scale to a substantial size:"

Realities – Plagiarism

- …and then there's this.

"Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size:"

"Bigtable is a conveyed stockpiling framework for overseeing organized information that is intended to scale to a substantial size:"

Agreement – Plagiarism

- You will be required to sign an agreement regarding plagiarism.

If you violate the agreement, you will receive a failing grade in the class.

- It's that simple.

Policies – Missing Things

- Missing class is not good. If you miss a lecture or a lab exercise, you will have to get that information and experience somewhere.
- On weekly assignments, there are no makeups. Some grading is automated, and I will discuss the assignments the following week. Get them done.
- On other assignments, except the final, if you are having trouble with a date, see me.

Policies – Class Conduct

- The usual rules about adult behavior apply.
- Kent is serious about academic honesty.
- Keep the laptop and phone distraction to a minimum when we're doing things together.
- I have no problem with snacks and drinks, and I will check on Kent's rules. This is a nice room, and was just remodeled, so keep it clean.
- If you need to leave between breaks, be discreet, please.

Policies – Extremes

- If something unusual is happening at the university, we might not have class if the university is closed.
- If something unusual happens to me, and I'm not here by 7:15, we won't have class.
- If either one of these happens, adjusted homework and lecture notes will be posted on Blackboard within 24 hours. *You will still be responsible for getting assignments done.*

Skills you will need

- Programming basics
  - How to use an IDE, how to debug
  - Basic language skills
    - SQL
    - Python
    - Javascript
    - Java
- Know your way around a Linux command line
- Some idea of how the web works
- Know how to use Git
- Some way to write scientific English

Stuff you need to get

- A Chrome Browser
  - On any OS
  - On a Chromebook or Chromebox
- A solid web connection
- Accounts on
  - PythonAnywhere.com (About $5/mo)
  - Codio.com (About $15/mo)

Stuff you need to read

- Downloadable/Web industry documentation
- Open source books as assigned
- Various system documentation
- On-line articles and industry commentary
- Research papers regarding big data topics

- You can learn from the original sources.

Class Project

Answer an Interesting Question

- Find a question worth answering
- Demonstrate the value of having an answer
- Locate some data
- Creating a working analysis environment
- Run the data load and analysis
- Produce the answer
- Evaluate the effectiveness of the effort
- Write a short paper about it

Online Services

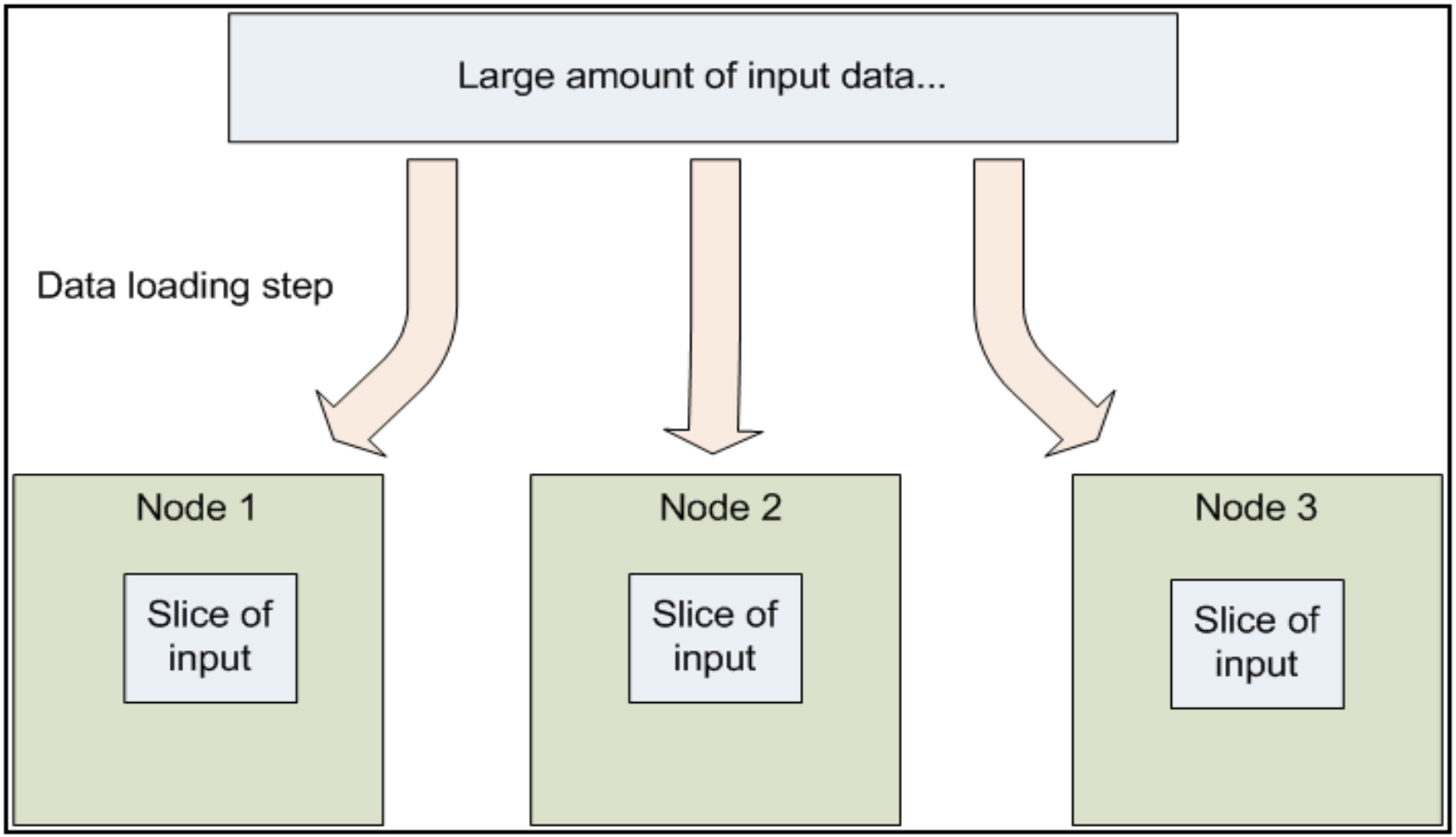
- IAAS – Codio.com
  - General purpose Unix boxes
  - Lots of capability
  - Throwaway boxes
- PAAS – PythonAnywhere.com
  - Hardened Unix Server
  - Production Web Capabilities
  - General purpose Python computation

Source Control

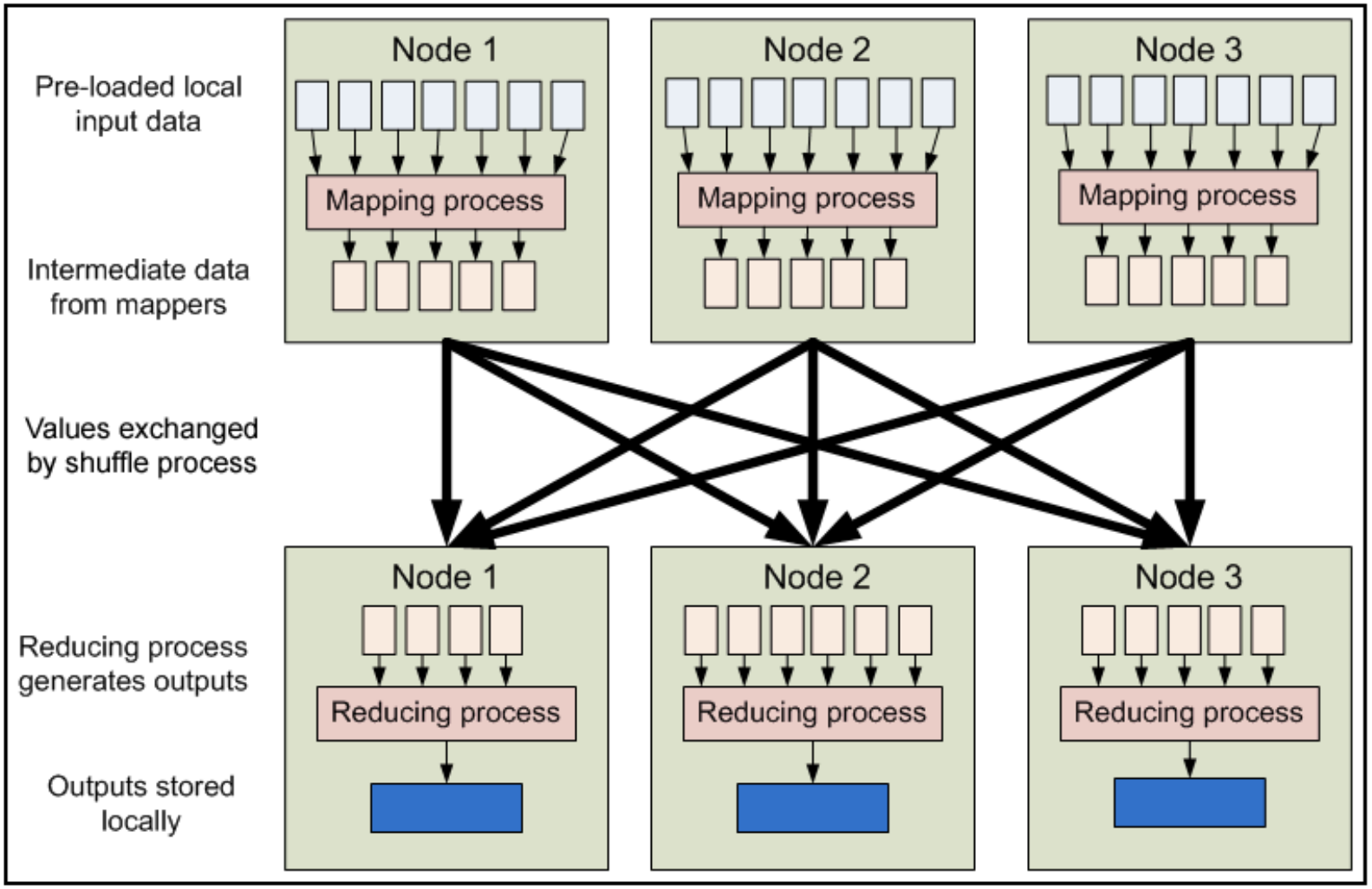
- GitHub
  - www.github.com
  - Get an account!
- Class content will be posted there.
  - www.github.com/gregdelozier/bigdata
  - …/lectures/<slides go here>
  - .../data/<datasets go here>
  - .../code/<various projects go here>
  - Clone this repo and use it.

Demo Time

Data Loading



Map Reduce


Pre-loaded local input data
Intermediate data from mappers
Values exchanged by shuffle process
Reducing process generates outputs
Outputs stored locally
Node 1
Node 2
Node 3
Mapping process
Mapping process
Mapping process
Node 1
Node 2
Node 3
Reducing process
Reducing process
Reducing process

Mapping

- Take a collection of key, value pairs
- Map onto a different collection of key, value pairs

Map(k1,v1) -> (k2,v2)

Shuffling

- (A,1),(B,2),(C,3) (B,3),(A,2),(C,1)
- (D,1),(C,1),(B,2) (A,5)

(Shuffle)

- (A,(1,2,5)) (B,(2,3,2)) (C,(3,1,1)) (D,(1))

Reduce

- Take a key and a list of values
- (A,(1,2,5)) (B,(2,3,2)) (C,(3,1,1)) (D,(1))
- Produce a smaller list of values

Reduce(k1,(v1,v2,v3)) -> (v4..)

- (A,8) (B,7) (C,5) (D,1)

Stream Text Pairs

- Imagine a collection of key, value pairs
- Express them on lines as
  - <key><tab><value>
- If <value> is missing, it's null

Stream Mapper

- Each mapper launches as a process
- For input
  - converts KV pairs to input lines
  - feeds lines to stdin
- For output
  - collects lines on stdout
  - converts output lines to KV pairs

Stream Python Mapper

```
#!/usr/bin/env python

import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    # increase counters
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        #
        # tab-delimited; the trivial word count is 1
        print '%s\t%s' % (word, 1)
```

Stream Reducer

- Each reducer launches as a process
- For input
  - converts KV pairs to input lines
  - feeds lines to stdin
- For output
  - collects lines on stdout
  - converts output lines to KV pairs
- Non-zero exit codes indicate failure

Stream Python Reducer

```python
#!/usr/bin/env python

from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # parse the input we got from mapper.py
    word, count = line.split('\t', 1)

    # convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

    # this IF-switch only works because Hadoop sorts map output
    # by key (here: word) before it is passed to the reducer
    if current_word == word:
        current_count += count
    else:
        if current_word:
            # write result to STDOUT
            print '%s\t%s' % (current_word, current_count)
        current_count = count
        current_word = word

# do not forget to output the last word if needed!
if current_word == word:
    print '%s\t%s' % (current_word, current_count)
```

Stream Preparation

- Make sure files are executable

  $ chmod +x filename.py

- Test scripts first with ordinary data
  - cat input/stream1/* | sort | wc
  - head -n1000 input/moby/moby.txt | ./mapper.py | sort | ./reducer.py

Demo Time

Homework – Due Next Week

- Get these examples running in your space.
  - Get the dataset
  - Write the stream programs
  - Demonstrate that they run
  - Submit the programs and the first page of output
    - Page is either the first 100 lines or the interesting stuff.
- Homework submission will be on Blackboard
  - (until I can find something better.)

Office Hours

- I will be setting up online office hours.
- Time suggestions?
- Instructions will be sent out.

Next Week

- I will be out of town
- I will be holding class online
- You'll need a browser with audio
- If you can't attend, get the content