CS15-319 / 15-619
Cloud Computing

Recitation 3
September 10th & 13th, 2013
Bugs!

• Checkpoint 1 grading
  – If you suspect a bug in a question, do let us know.
  – We have manually graded some questions in Checkpoint Quiz 1

• Encounter a general bug:
  – Post on Piazza

• Encounter a grading bug:
  – Post Privately on Piazza
UNIT 1: Checkpoint Quiz 1

- Stats
  - Average is 89%
  - Grades:
Unit 2: Data Centers

• Start reading first 2 modules:
  – Module 3: Data Center Trends
  – Module 4: Data Center Components
  – Module 5: Design Considerations
  – Unit 2: Checkpoint Quiz

<table>
<thead>
<tr>
<th>UNIT 2: Data Centers</th>
<th>Module 3: Data Center Trends</th>
<th>Module 4: Data Center Components</th>
<th>Module 5: Design Considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiz 2: Data Centers</td>
<td>Checkpoint</td>
<td>Not yet assigned</td>
<td>Due date TBD by instructor</td>
</tr>
</tbody>
</table>
Amazon Web Services (AWS) Account

- For students who just enrolled:
  - Create your AWS account
  - Send you account information to Jason Boles (jboles@cmu) with Email Subject: Request to add account to Consolidated Bill
Student Questions on Piazza

• Why was my script killed?
  – Running out of memory
  – Use larger instances? Maybe, maybe not.

Please ask public questions when possible.
Project 1 Student Progress

• Introduction to Big Data:
  – Sequential Analysis: Average is: **98%**
  – Elastic MapReduce: **this week**
Primer on MapReduce

• The idea of MapReduce

Please tell me how many times does the word “Apple” appear in these books?
Primer on MapReduce

• The idea of MapReduce

How many times does the word “Apple” appear in these books?

I heard 6 “Apple”s!
Primer on MapReduce

• The idea of MapReduce

How Do I know Who is the “Apple” Man?

No, you Don’t!
Primer on MapReduce

- The idea of MapReduce

Map Phase

- Orange, 1
- Blueberry, 1
- Blueberry, 1
- Apple, 1
- Apple, 1
- Apple, 1
- Orange, 1
- Apple, 1
- Apple, 1
- Orange, 1
- Blueberry, 1

Reduce Phase

Magic Box (Shuffle, sort, merge)

- Orange, 1
- Blueberry, 1
- Apple, 1
- Orange, 1
- Blueberry, 1

Reducer
Primer on MapReduce

- The idea of MapReduce

A full, simplified view of the phases, stages, tasks, data input, data output and data flow in the MapReduce analytics engine. (from Unit 4 Page 102)
Primer on MapReduce

- The idea of MapReduce

Map Phase

Orange, 1
Blueberry, 1
Blueberry, 1
Apple, 1

Apple, 1
Apple, 1
Apple, 1
Orange, 1

Apple, 1
Apple, 1
Orange, 1
Blueberry, 1

Reduce Phase

Black Box (Shuffle, sort, merge)

Orange ?
Apple ?
Blueberry ?
Primer on MapReduce

• Mapper
  – Input: lines in files in our project
  – Output: key-value pairs
    • Keys are used in Shuffling and Merge to find the Reducer that handles the intermediate output for that specific key. (in our example, Apple, Orange and Blueberry are keys)
    • Values are messages sent from mapper to reducer (in our case it is always 1)
    • Mappers’ output is intermediate because reducers will receive the key-value pairs and take them as input.
Primer on MapReduce

• Reducer
  – Input: **key-value pairs**
  – Output: the final result we need
    • Depends on what we want, our code should process the value in the key-value pairs that we got accordingly (in the word count example, we just add up all the values).
Primer on MapReduce

- In the projects of this course we are going to run MapReduce on 2 Platforms
  1. Amazon Elastic MapReduce (EMR)
  2. Hadoop (later projects)
Demos

• Wordcount in MapReduce
• S3 buckets and permissions
• Introduction to EMR
• Debugging EMR using logs
Discussions

• Your questions...
### Upcoming Deadlines

#### Project 1

**Introduction to Big Data Analysis**

<table>
<thead>
<tr>
<th>Task</th>
<th>Status</th>
<th>Due Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential Analysis</td>
<td>Checkpoint</td>
<td>Available Now</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Due 9/8/13 11:59 PM</td>
</tr>
<tr>
<td>Elastic MapReduce</td>
<td>Checkpoint</td>
<td>Available 9/9/13 12:01 AM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Due 9/15/13 11:59 PM</td>
</tr>
</tbody>
</table>

#### UNIT 2: Data Centers

**Module 3: Data Center Trends**

**Module 4: Data Center Components**

**Module 5: Design Considerations**

**Quiz 2: Data Centers**

<table>
<thead>
<tr>
<th>Task</th>
<th>Status</th>
<th>Due Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Checkpoint</td>
<td>Available 9/16/13 12:01 AM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Due 9/19/13 11:59 PM</td>
</tr>
</tbody>
</table>