15-319 / 15-619
Cloud Computing

Recitation 4

September 16th & 18th, 2014

Announcements

- Watch the videos in the project writeup
  - 1 video < 10 hours of debugging
- Starting this week (week 4) we will reduce handholding on Piazza
  - If you have not done enough work to investigate the reason of your issue we will not answer your question
  - Make Piazza more interesting, discuss new topics

Announcements

- In the code submission (zipped) file, do not include
  - Packages
  - Folder structures
  - Input or output data

- In the code submission (zipped) file, do include a ***readme*** text file with
  - Names and functionality of the files included
  - How to run your code
  - How to get the answers you submitted on OLI

Announcements

- Encounter a general bug:
  - Your responsibility to search first
  - Post on Piazza publicly
- Encounter a grading bug:
  - Post Privately on Piazza
- Do not post your code on Piazza
- Do not forget to Tag your instances
  - Key: Project Value: 2.1

Last Week's Reflection

- You have completed
  - Sequential Analysis
  - Elastic MapReduce

- You should have learned
  - Why MapReduce for big data
  - How MapReduce works
  - How to program Mapper & Reducer
  - Performance/cost tradeoff
  - How to narrow down bugs by using logs

Project 1.2 Checkpoint

- We will manually grade Question 1
  - Always make sure that your code is readable
    - Follow style presented in Recitation 2

Violation	Penalty of the project grade
Using any instance type in your cluster that has an on-demand pricing higher than m1.large	-10%
Using more than 20 core+task instances in your EMR cluster to run the job	-10%
Spending more than $15 for this project checkpoint	-10%
Failing to tag your instance for this project	-10%

Piazza Questions 1

- Late policy
  - We do not have a late policy!
    - All deadlines are hard. No exceptions.
    - No excuses, and no special cases are allowed.
  - Please start early!
    - We are working with a public cloud infrastructure, things are bound to take more time or break. One of the reasons why this experience is so valuable.

Piazza Questions 2

- Elastic MapReduce Billing Question
  - Normalized Hours (Elastic MapReduce)

Date	Elapsed Time	Normalized Instance Hours
1 11:59 EDT	1 hour 46 minutes	40

- 1 hour of m1.small = 1 hour normalized compute time
- 1 hour of m1.medium = 2 hours normalized compute time
- 1 hour of m1.large = 4 hours normalized compute time
- 1 hour of m1.xlarge = 8 hours normalized compute time
- 1 hour of c1.medium = 2 hours normalized compute time

Piazza Questions 2

- Elastic MapReduce Billing Question
  - EC2 cost + EMR cost
  - Elastic MapReduce Pricing (On demand)
  - for US East (N. Virginia)

General Purpose - Current Generation

m3.xlarge	$0.280 per Hour	$0.070 per Hour
m3.2xlarge	$0.560 per Hour	$0.140 per Hour

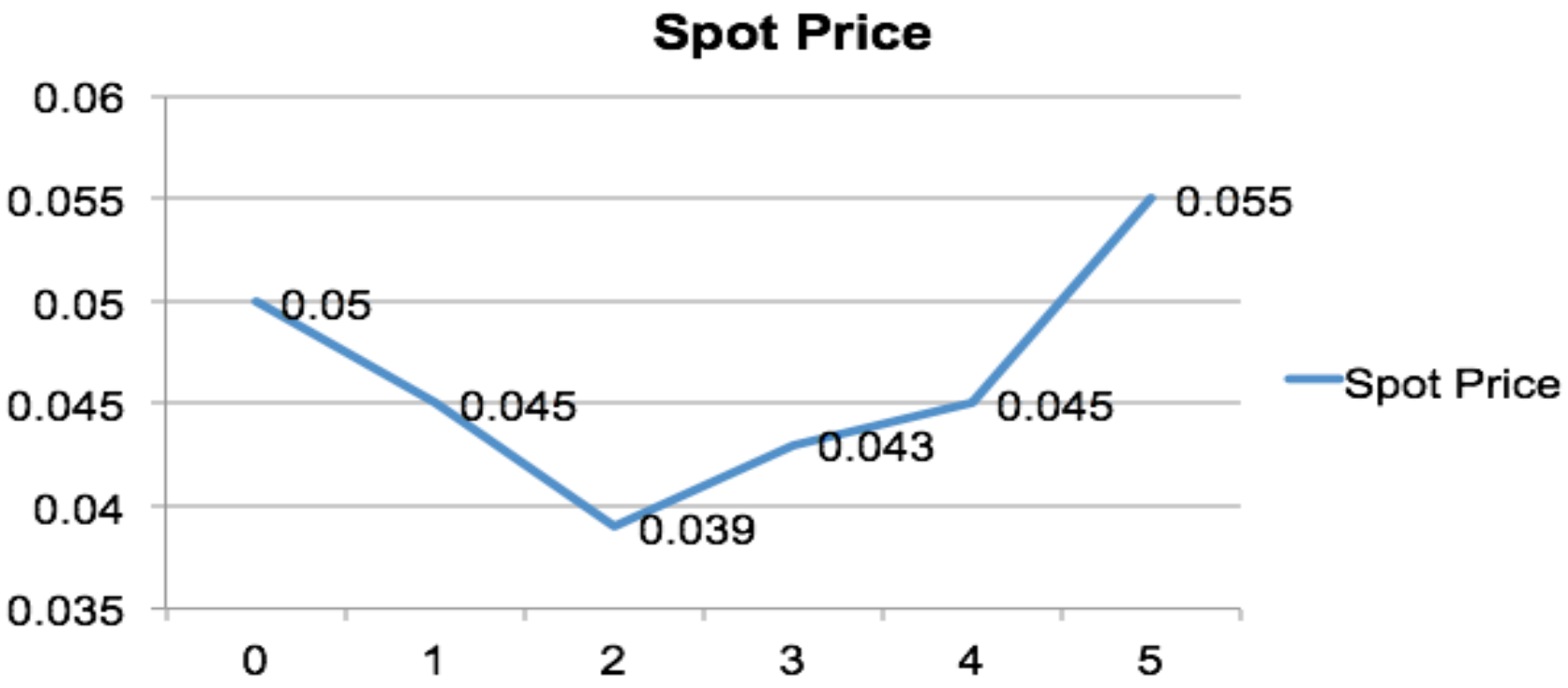
General Purpose - Previous Generation

m1.small	$0.044 per Hour	$0.011 per Hour
m1.medium	$0.087 per Hour	$0.022 per Hour
m1.large	$0.175 per Hour	$0.044 per Hour
m1.xlarge	$0.350 per Hour	$0.088 per Hour

Billing price = ($0.087 + $0.022) *2 instances * 2 hours = 0.436

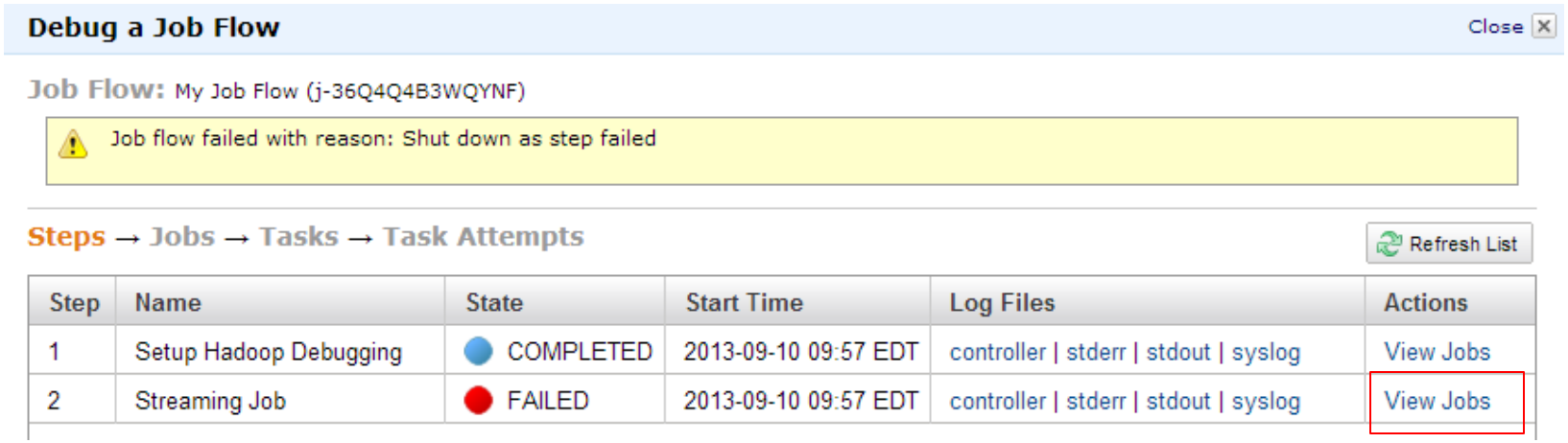
Piazza Questions 3

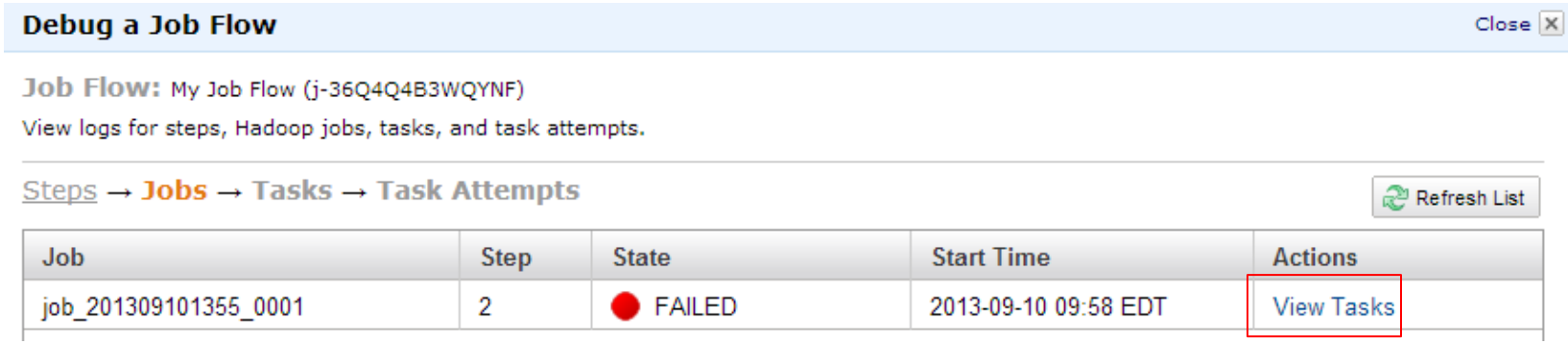
- Spot price (History $0.030~$0.060)



Piazza Questions 4

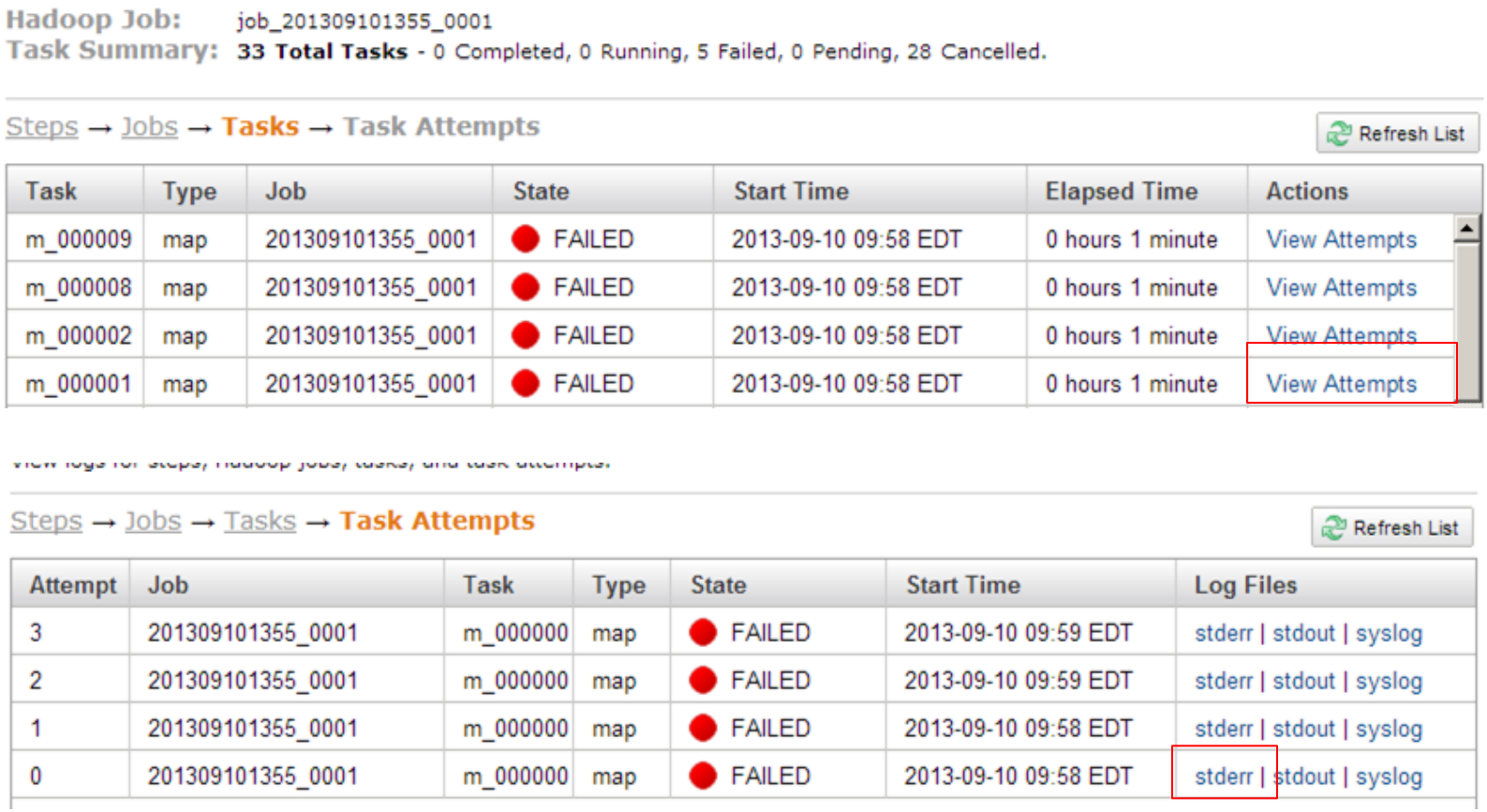
- Elastic MapReduce Debugging





Piazza Questions 4

- Elastic MapReduce Debugging



Hadoop Job: job_201309101355_0001
Task Summary: **33 Total Tasks** - 0 Completed, 0 Running, 5 Failed, 0 Pending, 28 Cancelled.

Steps → Jobs → **Tasks** → Task Attempts — Refresh List

Task	Type	Job	State	Start Time	Elapsed Time	Actions
m_000009	map	201309101355_0001	FAILED	2013-09-10 09:58 EDT	0 hours 1 minute	View Attempts
m_000008	map	201309101355_0001	FAILED	2013-09-10 09:58 EDT	0 hours 1 minute	View Attempts
m_000002	map	201309101355_0001	FAILED	2013-09-10 09:58 EDT	0 hours 1 minute	View Attempts
m_000001	map	201309101355_0001	FAILED	2013-09-10 09:58 EDT	0 hours 1 minute	View Attempts

Steps → Jobs → Tasks → **Task Attempts** — Refresh List

Attempt	Job	Task	Type	State	Start Time	Log Files
3	201309101355_0001	m_000000	map	FAILED	2013-09-10 09:59 EDT	stderr | stdout | syslog
2	201309101355_0001	m_000000	map	FAILED	2013-09-10 09:59 EDT	stderr | stdout | syslog
1	201309101355_0001	m_000000	map	FAILED	2013-09-10 09:58 EDT	stderr | stdout | syslog
0	201309101355_0001	m_000000	map	FAILED	2013-09-10 09:58 EDT	stderr | stdout | syslog

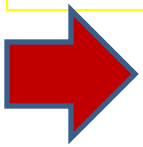
Piazza Questions 4

- Elastic MapReduce Debugging

```
/mnt/var/lib/hadoop/mapred/taskTracker/hadoop/jobcache/job_201309101355_0001/attempt_201309101355
/mnt/var/lib/hadoop/mapred/taskTracker/hadoop/jobcache/job_201309101355_0001/attempt_201309101355
unexpected token `('
/mnt/var/lib/hadoop/mapred/taskTracker/hadoop/jobcache/job_201309101355_0001/attempt_201309101355
is_in_filtered_title(line_item_list):'
java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2
        at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:372)
        at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:586)
        at org.apache.hadoop.streaming.PipeMapper.map(PipeMapper.java:125)
        at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:50)
        at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:36)
```

What happens this Week

- Unit 2: Data Centers
  - Module 3: Data Center Trends
  - Module 4: Data Center Components
- Read and complete:
  - Module 5: Design Considerations
  - Unit 2: Checkpoint Quiz
    - 150 minutes, due September 18th, 2014, Pittsburgh



UNIT 2: Data Centers			
	Module 3: Data Center Trends		
	Module 4: Data Center Components		
	Module 5: Design Considerations		
	Quiz 2: Data Centers	Checkpoint	Available 15/09/14 12:01 AM Due 18/09/14 11:59 PM

Module 5: Design Considerations

- Design considerations
  - Requirements and Geographic Location
  - Costs
  - Power and Efficiency
  - Redundancy

- Challenges and Requirements
  - Scalability
  - Network Topologies
  - Utilizations & Resiliency
  - Security

Amazon data center

Quiz 2

- Quiz 2 is open
  - This assignment is timed. You will have 150 minutes to complete the attempt once you begin.
  - Deadline for completion is Sep 18, 11:59 PM Pittsburgh
  - Late submissions are NOT accepted
  - You may not start the assignment after the deadline has passed.
  - Maintain your own timer from when you start the quiz.
- You only have 1 attempt
- Question types
  - multiple choice, fill-in-the-blanks, numeric questions
    - most likely but not limited to these
- Feedback on Quiz 2 is released after the deadline passes

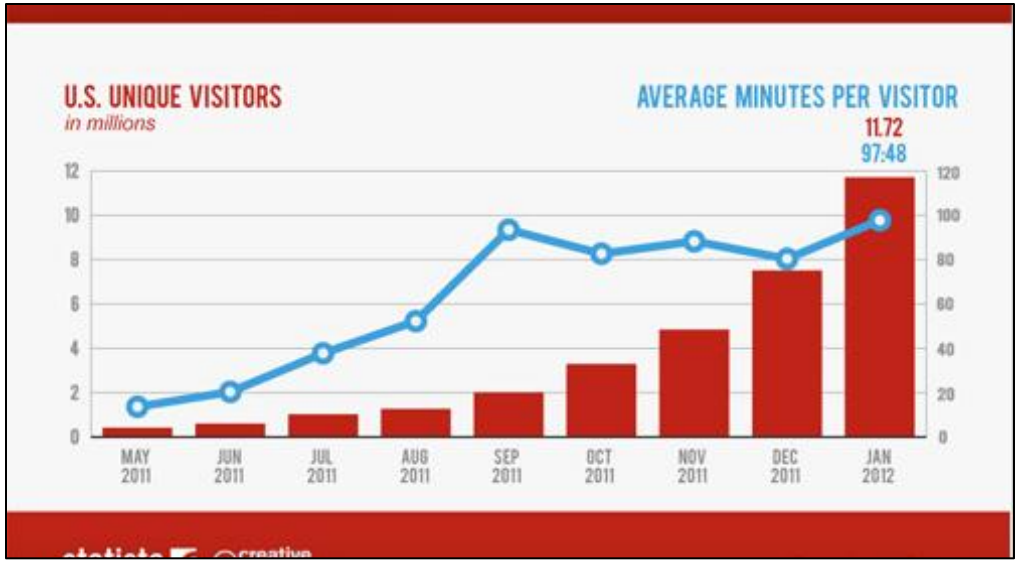
This Week's Project

- Project 2.1: Introduction to APIs
  - MSB Recruitment Exam
- Start early!
  - Project 2.1, due 21th September, (Pittsburgh time)
  - You have three attempts, but no late submission

Project 2			
	Introduction and APIs		
	MSB Recruitment Exam	Checkpoint	Available Now Due 09/21/14 11:59 PM

Pinterest

http://www.pinterest.com/



- Started with 12 engineers, now 300+ employees
- Monthly active users: 40 million
- Growth in web traffic from 9/12-9/13: 66.52%

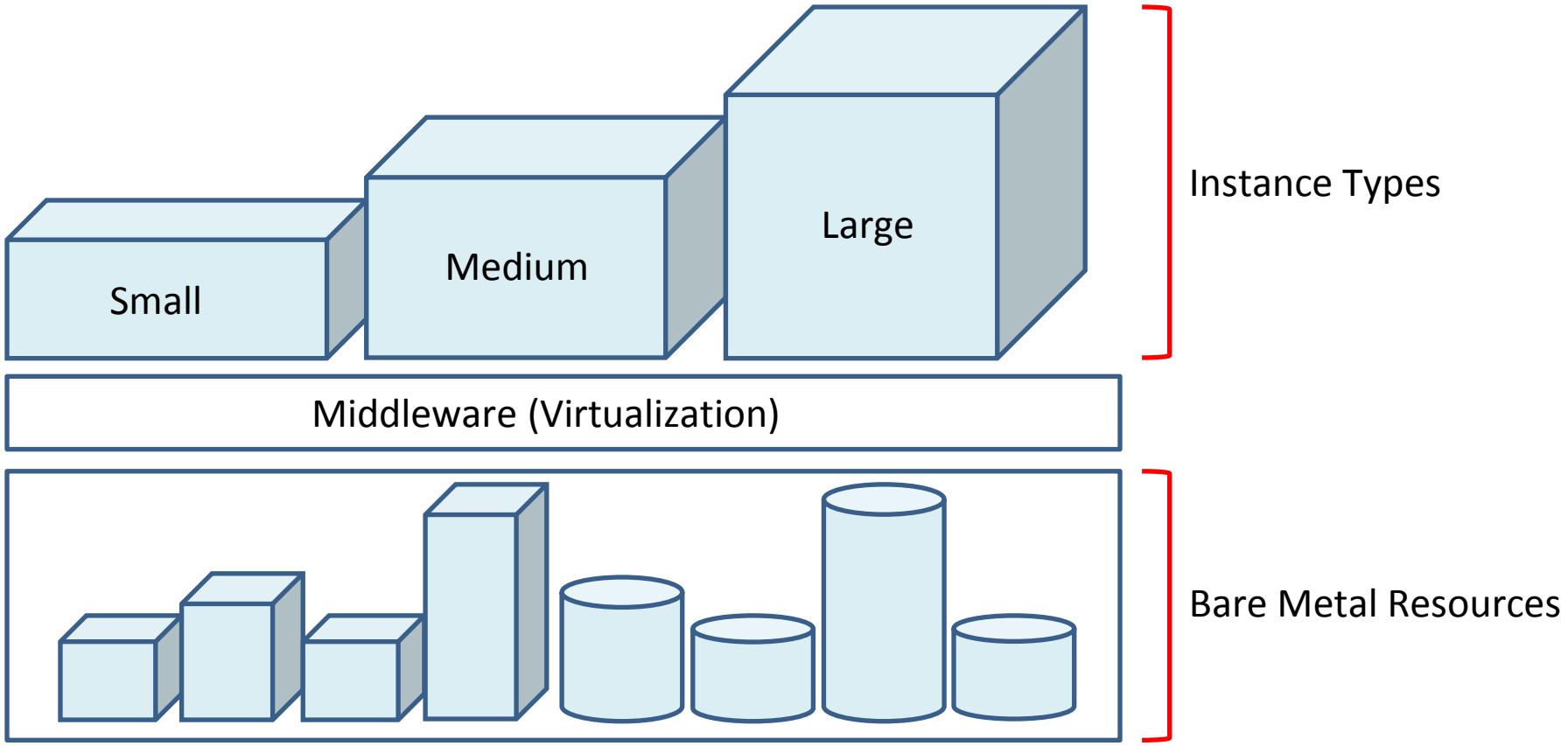
In Unit 1, we know that cloud computing provides several advantages, including:

- Elasticity
- Reduced upfront cost
- Reduced maintenance cost

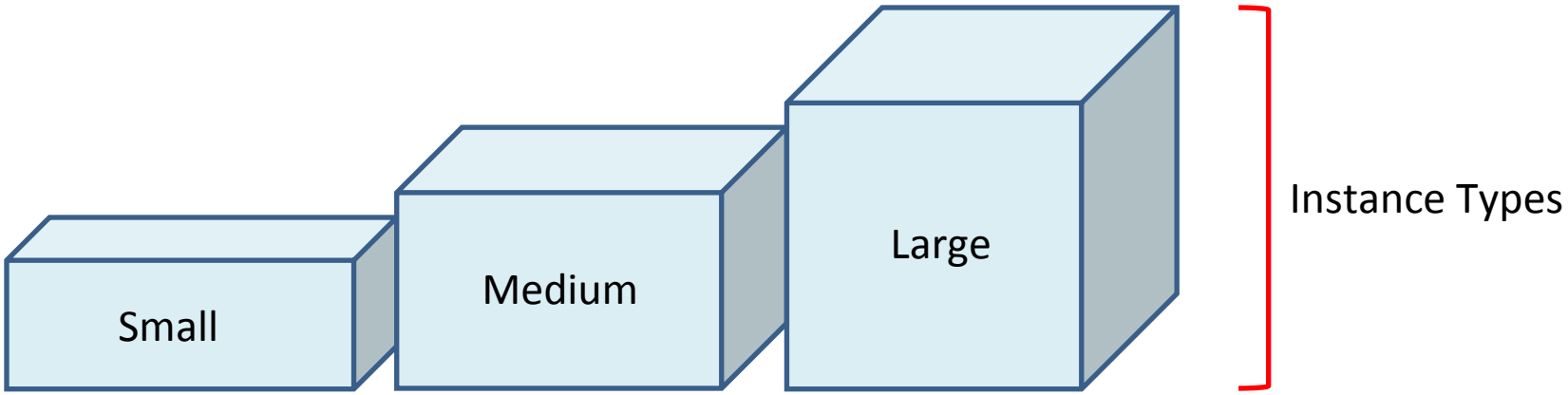
Pinterest

- Utilizes AWS
  - S3
    - File storage
  - Elastic MapReduce
    - Data analysis
  - Auto Scaling
    - Scale up and down
  - Elastic LoadBalancer
    - Distribute traffic Data analysis

Resources in Cloud Infrastructure



Maximize Requests per Dollar

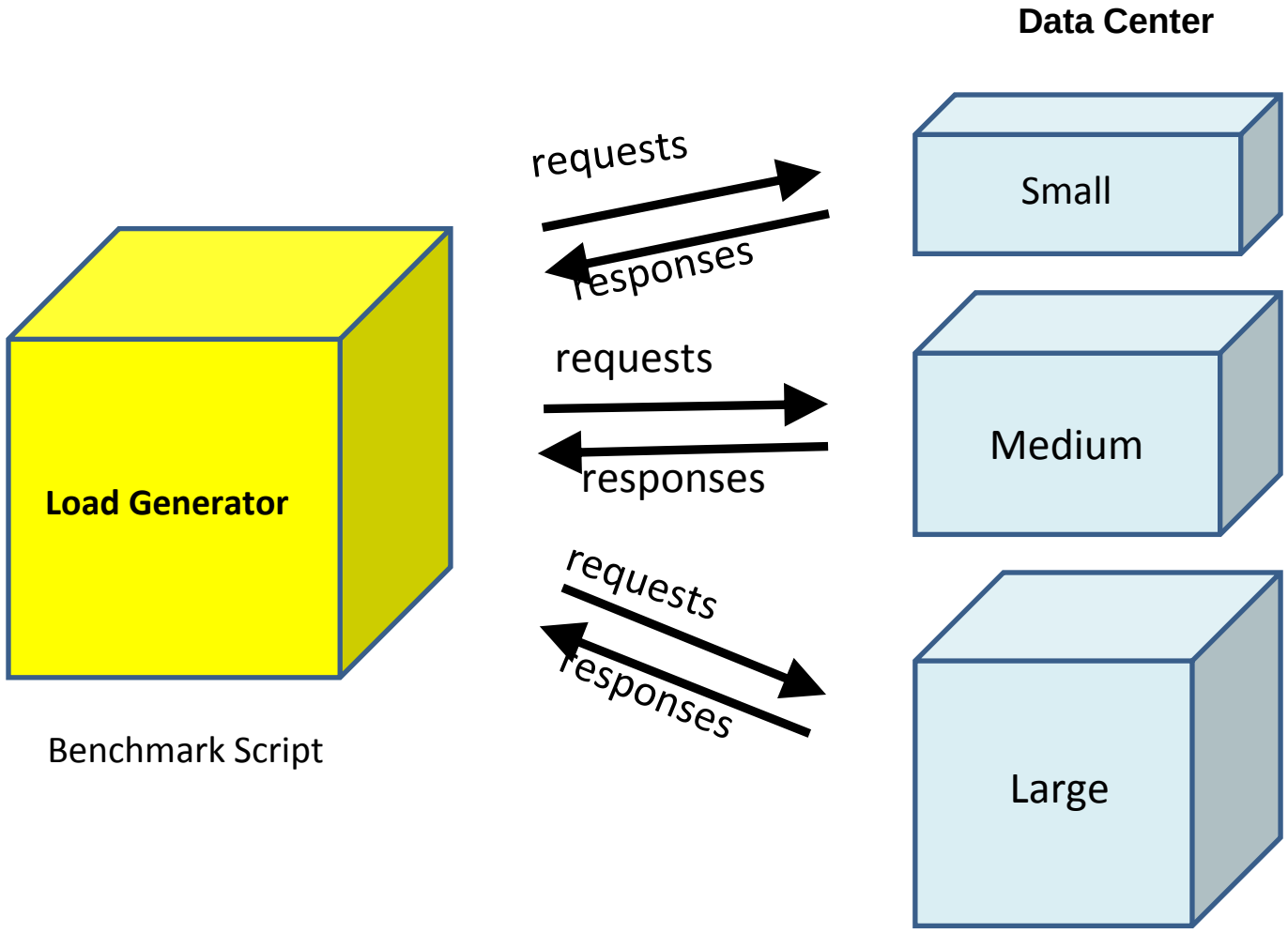


X REQUESTS ***Y*** REQUESTS ***Z*** REQUESTS

*COST **A*** < *COST **B*** < *COST **C***

GOAL: MAXIMIZE (REQUESTS/DOLLAR)

Load Testing Request & Response Flow



Amazon APIs

- Supported APIs
  - Command Line Interface API Tools
  - AWS SDK for Java
  - AWS SDK for Python

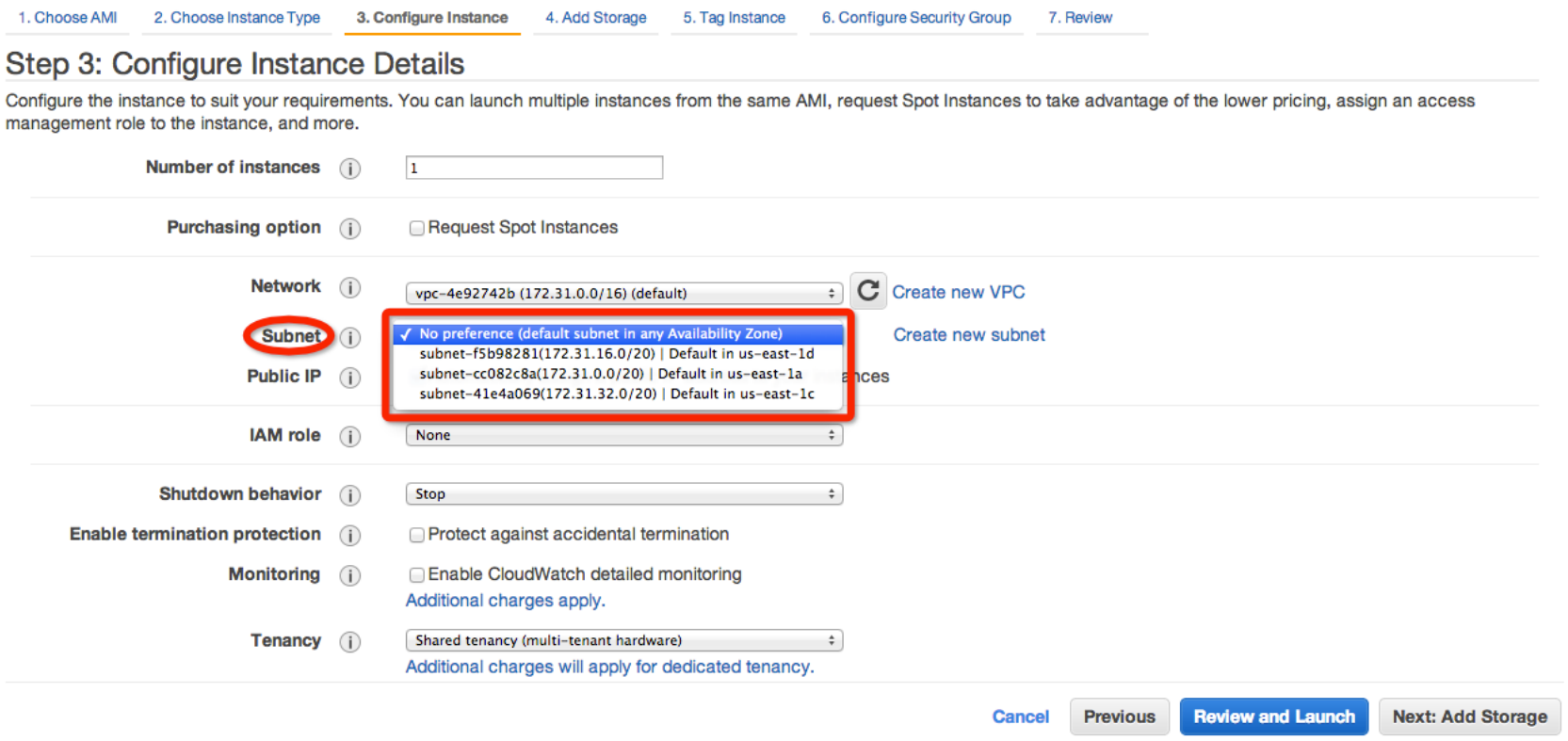
Highlight of This Week's Project

- Manually navigate through all major components. Some of the checkpoint quiz questions may rely on this part
- Install the API based on your choice
  - Follow AWS instructions and ours
- Understand how to use the API and complete the programmatic automation of scaling up the servers to handle network traffic.
  - Checkpoint quiz questions associated with this part

Note: You can always read the checkpoint quiz questions in advance to know what you should be aware of.

Other Reminders

- Make sure the Load Generator and Data Center VMs are in the same subnet (availability zone)



Other Reminders (cont')

- Terminate instances vs. Stop instances
  - Stop will still charge for VM storage (EBS volumes)
    - Stop is a good idea when you need a break
- DO NOT add your credentials in your code
- S3 URL Validation and Naming Convention
  - Always validate your link before submitting the link
    - http://ec2-54-225-106-182.compute-1.amazonaws.com/
  - Follow instructions

Penalties

- If
  - No tag ➔ **10%** penalty
  - Expenditure > project budget ➔ **10%** penalty
  - Expenditure > project budget * 2 ➔ **100%** penalty
  - Copy any code segment from ➔ lowest penalty is **200%**
    - Other students
    - Previous students
    - Internet (e.g. Stackoverflow)
- Do not work on code together
  - This is about learning
  - If you do, we *will* find out and take action

Submitting Code

- Submit your code on S3, by the deadline
  - Submitting the wrong S3 URL on OLI will be penalized
  - Submitting an incorrectly configured bucket will be penalized (see the instructions in the Project Guidelines page in the Project Primer on OLI)

- We will manually grade all code
  - Be sure to make your code readable
    - Preface each function with a header that describes what it does
    - Use whitespace well.
      Indent when using loops or conditional statements
    - Keep each line length to under 80 characters
    - Use descriptive variable names
    - For more detail, please refer to www.cs.cmu.edu/~213/codeStyle.html
  - If your code is not well documented and is not readable, we will deduct points
    - Documentation shows us that you know what your code does!
    - The idea is also NOT to comment every line of code

S3 Code Submission Guidelines

- Make your submission a single zip file (.zip) with name "**project<no>_AndrewID_q<no>"**.
- Pack all your code, in ".java", ".py", ".rb", ".sh" or other formats in the zip file.
- Do **NOT** submit associated libraries and binary files (.jar and .class files).
- Please do **NOT** submit multiple s3 URLs in the text field on OLI.
- Create a single submission bucket for all of your code, using the folder hierarchy illustrated in the project primer on page 151 on OLI.
- Do NOT submit AWS Credential Files (aws.credentials) or any other files that contain your AWS Keys within your bucket.

  ➔ **10%** penalty
- Do not make your buckets public.

Upcoming Deadlines

- Quiz 2: Data Centers
  - –Quiz Available Now
  - – Due 9/18/2014 11:59 PM EST
- Project 2: Introduction and APIs
  - – MSB Recruitment Exam
  - – Checkpoint Available Now
  - – Due 9/21/2014 11:59 PM EST
- •Arrange your time wisely at TOC week!