15-319 / 15-619
Cloud Computing

Recitation 10

March 28, 2017

Overview

- **Last week's reflection**
  - Project 3.3
  - Unit 4 - Module 15
  - Quiz 8

- **This week's schedule**
  - Unit 4 - Modules 16, 17
  - Quiz 9
  - Team Project Phase 1, due on this Sunday!

Project 3.3 Reflection

- You've explored
  - Sharding and Replication
  - Multithreaded programming
  - Strong Consistency model
    - Use AHEAD to keep proper order on all datastores
  - Bonus Task: Eventual Consistency
    - Non-blocking calls in the main thread

Project 3.3 Survey

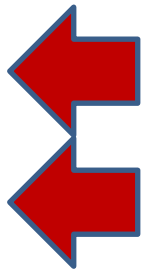
- Please complete the Project 3.3 Survey:

  https://piazza.com/class/iwo7h5zi9h96fw?cid=1216

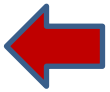
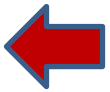
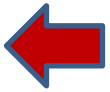
- Your feedback is very valuable!

Modules to Read

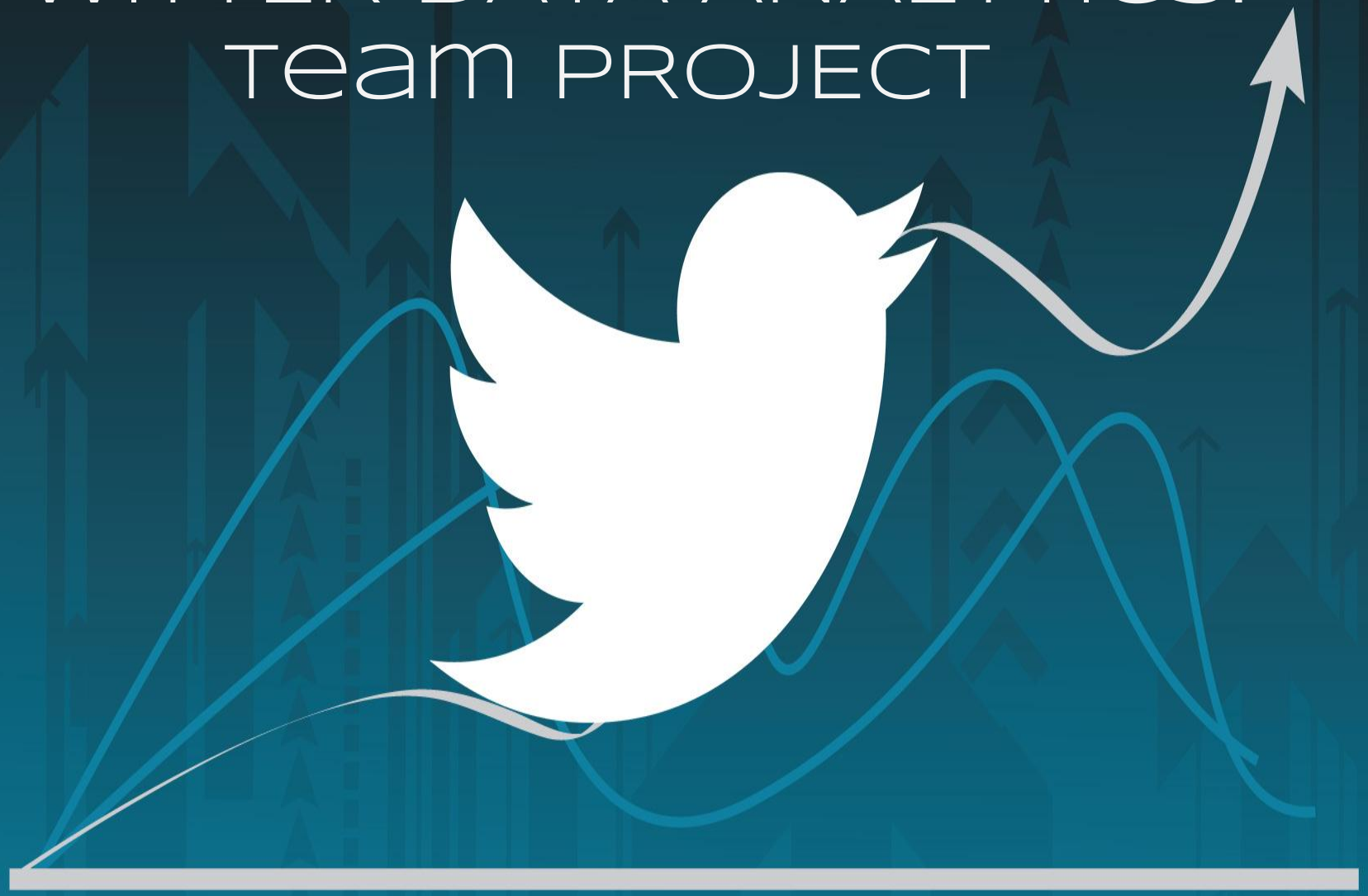
- UNIT 4: Cloud Storage
  - Module 14: Cloud Storage
  - Module 15: Case Studies: Distributed File System
    - HDFS
    - Ceph
  - Module 16: Case Studies: NoSQL Databases
  - Module 17: Case Studies: Cloud Object Storage



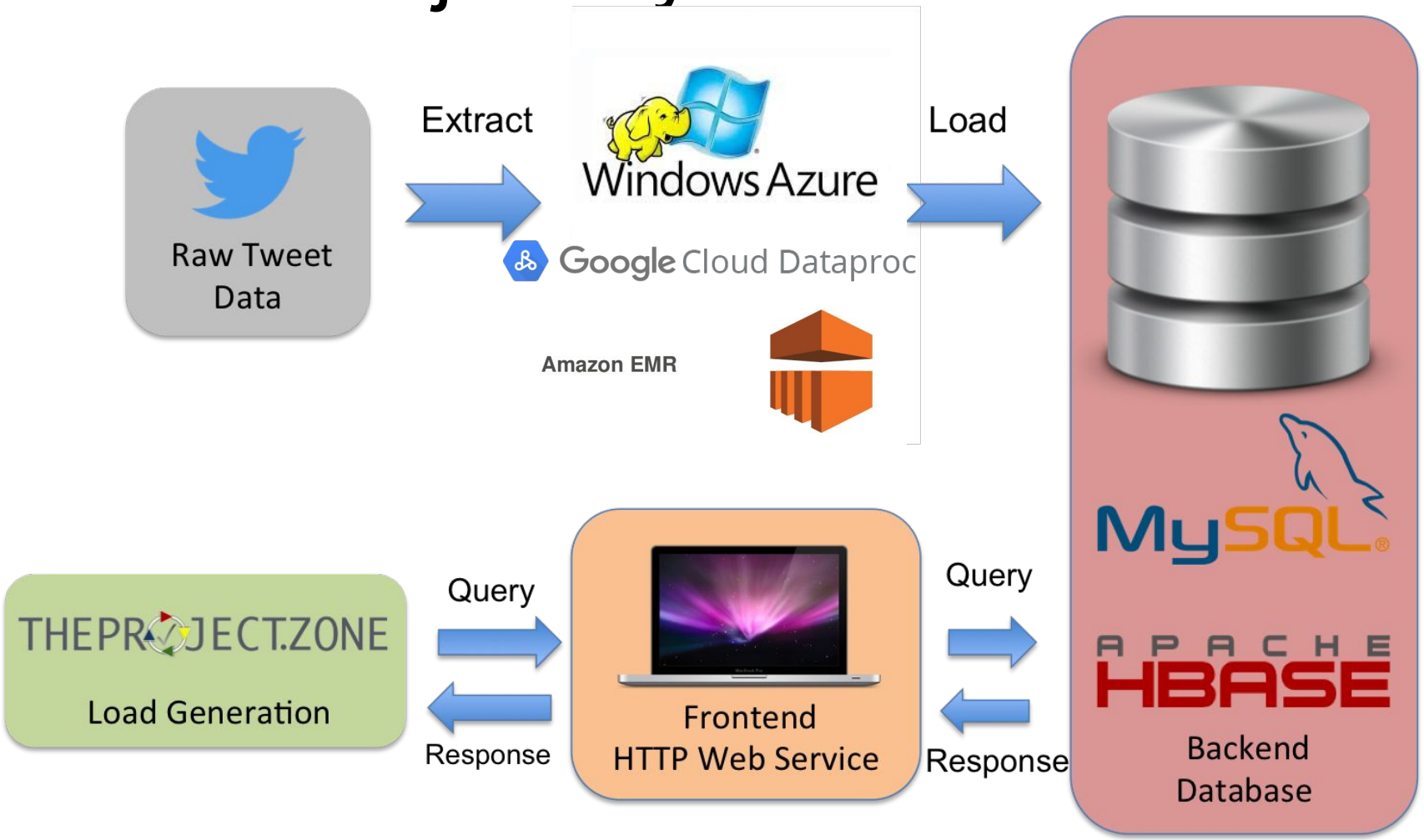
Upcoming Deadlines

- Quiz 9: Unit 5 - Module 16, 17
  - Due: 03/31/2017 11:59 PM Pittsburgh 
- Team Project: Q1, Q2M & Q2H, Q3M & Q3H
  - Phase 1 due on 04/02/2017 11:59 PM Pittsburgh 
- Team Project: Phase 1 code & report
  - Due on 04/04/2017 11:59 PM Pittsburgh 

TWITTER DATA ANALYTICS: Team PROJECT



Team Project System Architecture

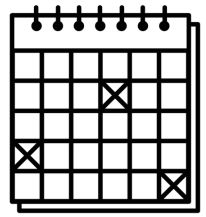


- Web server architectures
- Dealing with large scale real world tweet data
- HBase and MySQL optimization

Team Project

- Phase 1:
  - Q1
  - Q2 & Q3 (MySQL **AND** HBase)
- Phase 2
  - Q1
  - Q2 & Q3 & Q4 (MySQL **AND** HBase)
  - Live test
- Phase 3
  - Q1
  - Q2 & Q3 & Q4 & Q5 (MySQL **OR** HBase)
  - Live test

Team Project Time Table



Phase	Query	Start	Deadline	Code and Report Due
Phase 1	Q1	Monday 2/27/2017 00:00:01 EST	Sunday 3/12/2017 23:59:59 EST	-
	Q2 & Q3	Monday 2/27/2017 00:00:01 EST	Sunday 4/2/2017 23:59:59 EST	Tuesday 4/4/2017 23:59:59 EST
Phase 2	Q1, Q2, Q3, Q4	Monday 4/3/2017 00:00:01 EST	Sunday 4/16/2017 15:59:59 EST	-
	Live Test Q1, Q2, Q3, Q4	Sunday 4/16/2017 18:00:01 ET	Sunday 4/16/2017 23:59:59 EST	Tuesday 4/18/2017 23:59:59 EST
Phase 3	Q1, Q2, Q3, Q4, Q5	Monday 4/17/2017 00:00:01 ET	Sunday 4/30/2017 15:59:59 EST	-
	Live Test Q1, Q2, Q3, Q4, Q5	Sunday 4/30/2017 18:00:01 ET	Sunday 4/30/2017 23:59:59 EST	Tuesday 5/2/2017 23:59:59 EST

Note:

- There will be a report due at the end of each phase, where you are expected to discuss optimizations
- **WARNING: Check your AWS instance limits on the new account (should be > 10 instances)**

Team Project Phase 1

- Q1
  - Building a heartbeat and authentication web service.

- Q2 & Q3
  - Handling complex read-only queries.
  - Doing ETLs, building, configuring and optimizing Web Tier and Database Tier.
  - Test MySQL and HBase respectively.

Team Project Phase 1 Scoreboard

Top 10 Teams:

cc is my god	**75**
HongKongJournalists	**45**
CC don't play me	**45**
CCMaster	**45**
IN CC WE TRUST	**45**
MaLaoShiWanSui	**45**
Plus1s	**45**
NoBug	**45**
Red Cliff	**45**
Fregatidae	**45**

Good job!

Phase 1, Query 1 Reflection

- All teams did an excellent job!
  - Compared and tested different web frameworks
  - Choose and configured a fast frontend
- Remember to write your auto-deployment (orchestration) script for Query 1!

Phase 1, Query 2 Tips

- Use regex “\p{L}+” to match words in Java, **FOR Q2 ONLY.**
  - So that we only match the unicode letters. Simply ignore the unicode marks, i.e. “\p{M}”.
- If one hashtag appears in a tweet multiple times, the tweet should be weighted based on the frequency of this hashtag showing up.
  - Word frequency in this tweet should be calculated multiple times if it contains the hashtag more than once.
- Treat all contents in the `text` field the same, including hashtag. Meaning if the text contains hashtags, these hashtags can be considered as a word and need to be included in the keyword frequency calculation.
- Only hashtags in the request should be considered as case sensitive.

Phase 1, Query 3 FAQs

Question 1: How to calculate the topic score?

- Calculate the IDF score idf(w) of word w in the given range of tweets.
- Calculate the term frequency of word w in i-th tweet Ti.
- The impact score of i-th tweet Ti is impact(i).
- Topic score for word w is
  - SUM(Ti*idf(w)*ln(impact(i)+1)) (For tweets in given range)

Phase 1, Query 3 FAQs

Question 2: When to censor? When to exclude stop words?

- Censor in the Web Tier or during ETL. It is your own choice.
  - If you censor in ETL, consider the problem it brings to calculating the topic scores (two different words might look the same after censoring).

- You should count stop words when counting the total words for each tweet in order to calculate the topic score. Exclude stop words when calculating the impact score and selecting topic words.

Performance Tuning Tips

- To do performance tuning, you first need to identify which part of your system is the bottleneck.
  - Do profiling and monitoring on your system
    - Write a LG yourself to test your system performance
    - Use CloudWatch for resource utilization, etc.
  - Blindly trying different things is time-consuming, and doesn't work well (Amdahl's Law).

Performance Tuning Tips

- Think about the architecture of your system and what advantages and disadvantages it has compared to other settings
  - Sharding vs. Replication
  - ELB vs. Customized LB
  - Doing the calculation in Web Tier vs. in ETL, etc.

Performance Tuning Tips

- Web Tier
  - Are you using worker threads with an event-driven framework? (e.g. Using “blocking” in Vert.X)
  - Did you put too much computation at the Web Tier?
  - If you have multiple Web Tier servers, is the workload distributed evenly?
  - Have you optimized your algorithm? etc.

Performance Tuning Tips

- Database Tier
  - Try to reduce the number of rows / the size of data retrieved in each request.
  - Remember that Q2 & Q3 are read-only. You can choose schemas that are specifically optimized for Q2 & Q3.

Performance Tuning Tips

- Database Tier - MySQL
  - Tune the parameters
    - Check the official documentation
    - Google for MySQL performance tuning

- Database Tier - HBase
  - Tune the parameters
  - Be aware of data distribution and try to identify hotspots
  - Scan can be really slow, try to avoid it when possible
    - If not, try to scan as few rows as possible

Performance Tuning Tips

- Review what we have learned in previous project modules
  - Scaling out
  - Load balancing
  - Replication and Sharding
- Ask on Piazza or go to office hour if you are stuck for too long!

Reminder

- Your team has a total AWS budget of $50 for Phase 1
- Your web service should cost ≤ $0.88/hour, including:
  - EC2
    - We evaluate your cost using the On-Demand Pricing towards **$0.88/hour** even if you use spot instances.
  - EBS & ELB
  - Ignore data transfer and EMR cost
- Targets:
  - Query 2 - 7000 rps (for both MySQL and HBase)
  - Query 3 - 1500 rps (for both MySQL and HBase)

Upcoming Deadlines

- Quiz 9: Unit 4 - Module 16, 17

  Due: Friday 3/31/2017 11:59 PM EST

- Team Project: Phase 1

  Due: 4/2/2017 11:59 PM EST

- Team Project: Phase 1 code & report

  Due: 4/4/2017 11:59 PM EST

Questions?