

15-319 / 15-619

Cloud Computing

Recitation 11

Nov 7th 2017

Overview

- **Last week's reflection**
 - Project 4.1
 - Project 2.3
 - Unit 4 - Modules 16 & 17
 - Quiz 9
 - Team Project, Phase 1, report
- **This week's schedule**
 - Team Project, Phase 2, Queries, 1, 2, 3
 - Live test!
 - Unit 5 - Module 18
 - Quiz 10 (Due on **Thursday, 11/9**)
- **Twitter Analytics: The Team Project**

Reminders

- Monitor AWS expenses regularly and tag all resources
 - Check your bill both on AWS and TPZ
- Piazza Guidelines
 - If you need us to debug a specific submission, please give your submission ID
 - If you have a grading issue, please provide your andrew ID
- Utilize Office Hours
 - Take full advantage of the office hours

Project 2.3 Reflection

- Thanks for attempting a new project and providing helpful feedback!
- For you who completed P2.3, this project's grade will replace the lowest project score from P1.1, P1.2, P2.1, P2.2, P3.1, or P3.2
- You gained experience with FFmpeg, AWS Lambda, Azure Functions, GCP Cloud Functions, Rekognition, and CloudSearch
- Students had a bit of trouble with Task 2 but enjoyed using functions to build a real-world video indexing system

Project 4.1 Reflection

Takeaways from P4.1:

- Use the MapReduce Model and think like mappers and reducers. Understand the complete workflow and configurations to design and implement MapReduce applications.
- Use the UI to find logs and locate problems; and solve any dependency issues.
- Some bugs might look like memory or YARN scheduling issues, but they could be bugs or inefficiencies in your program! E.g. If you sort all the probabilities in your reducer, it will cause a reducer timeout.
- Optimize your program by using some advanced features of the MapReduce Framework such as Partitioner and Input/OutputFormat.

Phase 1 Report Feedback

- Good job for most teams
- The report helps you think about
 - The starting point for optimization: what to try, where is the bottleneck in our current implementation
 - How to further improve Q1 and Q2, and guide you to think about Q3
- If you explored enough for each question, you should have a lot to say in report.
- Hope you will have more “fancy tricks” or optimizations to share in future reports!!
 - Work hard to become one of the top teams!!!

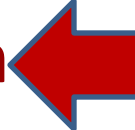
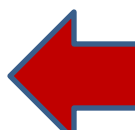

Modules to Read

- UNIT 5: Distributed Programming and Analytics Engines for the Cloud
 - Module 18: Introduction to Distributed Programming for the Cloud



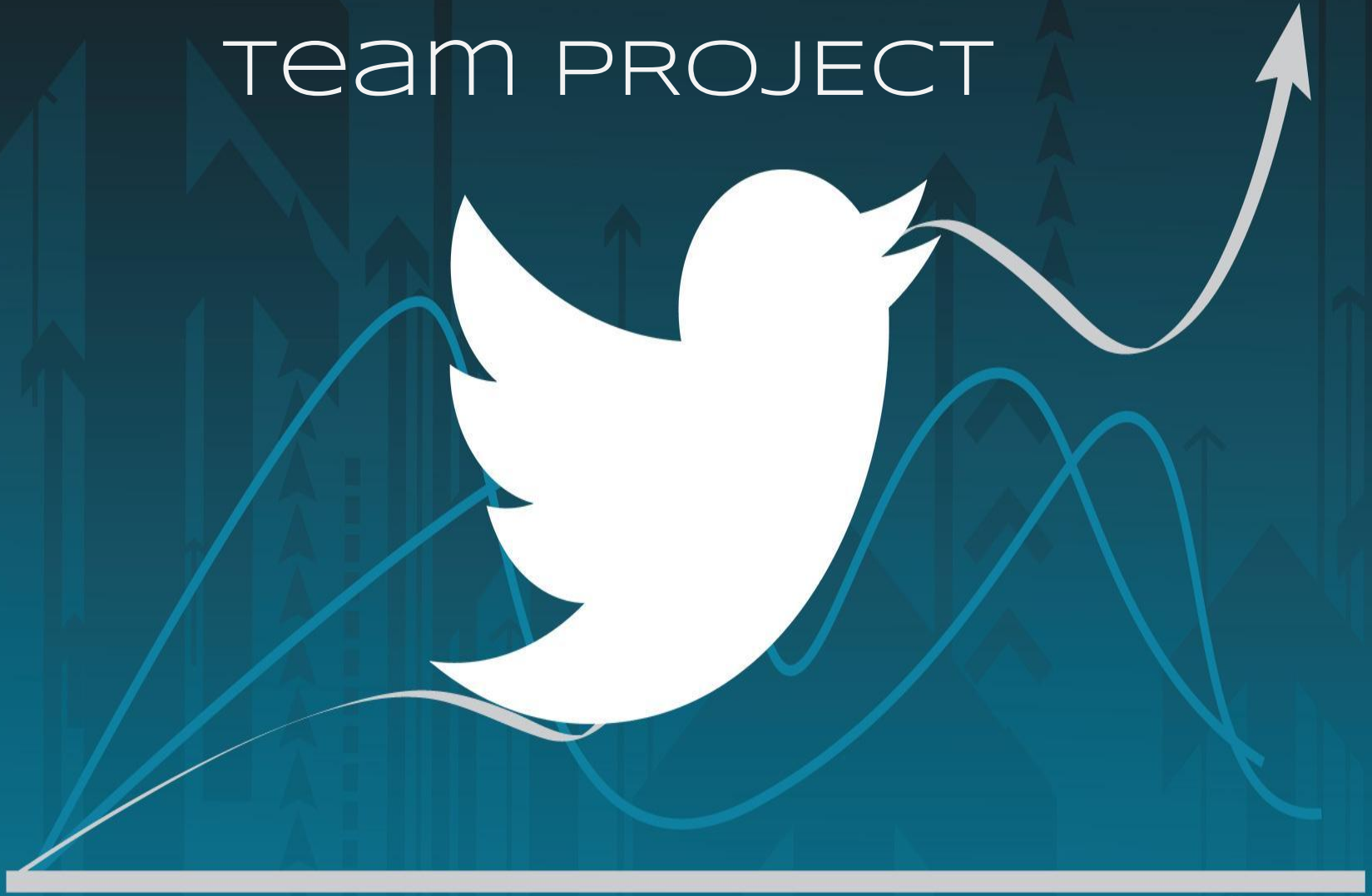
Upcoming Deadlines



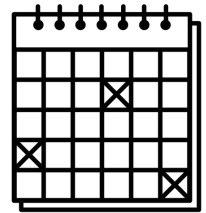
- Quiz 10: Unit 5 - Module 18
Due: Thursday, Nov 9th, 2017 11:59PM Pittsburgh 
- Team Project - Phase 2
Due: Sunday, Nov 12th 2017 3:59PM Pittsburgh 
- Team Project - Phase 2 - Live Test
Sunday, Nov 12th 2017 4PM-11PM Pittsburgh 

Questions?

TWITTER DATA ANALYTICS: Team PROJECT



Team Project Time Table



Phase	Query	Start	Deadline	Code and Report Due
Phase 1	Q1	Monday 10/9/2017 00:00:01 EST	Sunday 10/22/2017 23:59:59 EST	-
	Q2	Monday 10/9/2017 00:00:01 EST	Sunday 10/29/2017 23:59:59 EST	Tuesday 10/31/2017 23:59:59 EST
Phase 2	Q1, Q2, Q3	Monday 10/30/2017 00:00:01 EST	Sunday 11/12/2017 15:59:59 EST	-
	Live Test Q1, Q2, Q3	Sunday 11/12/2017 18:00:01 ET	Sunday 11/12/2017 23:59:59 EST	Tuesday 11/14/2017 23:59:59 EST
Phase 3	Q1, Q2, Q3, Q4	Monday 11/13/2017 00:00:01 ET	Sunday 11/26/2017 15:59:59 EST	-
	Live Test Q1, Q2, Q3, Q4	Sunday 11/26/2017 18:00:01 ET	Sunday 11/26/2017 23:59:59 EST	Tuesday 11/28/2017 23:59:59 EST

Team Project Phases 1&2 Review

- Q1
 - Building a heartbeat and authentication web service.
- Q2
 - Handling complex read-only queries.
 - Doing ETL, building, configuring and optimizing both the Web Tier and Database Tier.
 - Explore both MySQL and HBase.
- Q3
 - Handling range read requests
 - Try more optimizations in Web/DB tier, explore
 - schema, DB configurations, explore optimizations given the type of the query

Phase 1, Query 2 Tips

- Use regex “\p{L}+” to match words in Java
 - So that we only match the unicode letters.
- 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 as a keyword.
- Keywords and hashtags are case insensitive.

Query 3

- Problem Statement
 - In a time range and a user id range, which tweets have the most impact and what are the topic words?
- Impact score and topic words (see the write up for details)
 - Impact of tweets: Which tweet is “important”? Calculate using the effective word count and statistics like retweet count and follower count.
 - Topic words: In this given range, what words could be viewed as a “topic”? Done using TF-IDF.
- Request/Response Format
 - Request: Time range, uid range, #words, #tweets
 - Response: List of topic words with their topic score, as well as a list of censored tweets

Phase 2, Query 3 FAQs

Question 1: How to calculate the topic score?

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

Phase 2, 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 word 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 such as CPU, Network, Disk, etc.

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
 - 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?

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
 - Search 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 hours if you are stuck for too long!

Reminder

- Your team has a total AWS budget of **\$50** for Phase 2
- Your web service should cost \leq **\$0.83/hour**, including:
 - EC2
 - We evaluate your cost using the [On-Demand Pricing](#) towards **\$0.83/hour** even if you use spot instances.
 - EBS & ELB
 - Ignore data transfer and EMR cost
- Live Test Targets:
 - Query 1 - 28000 rps
 - Query 2 - 10000 rps (for both MySQL and HBase)
 - Query 3 - 1500 rps (for both MySQL and HBase)

Reminder, Live Test this Sunday 11/12

Phase	Query	Start	Deadline	Code and Report Due
Phase 1	Q1	Monday 10/9/2017 00:00:01 EST	Sunday 10/22/2017 23:59:59 EST	-
	Q2	Monday 10/9/2017 00:00:01 EST	Sunday 10/29/2017 23:59:59 EST	Tuesday 10/31/2017 23:59:59 EST
Phase 2	Q1, Q2, Q3	Monday 10/30/2017 00:00:01 EST	Sunday 11/12/2017 15:59:59 EST	-
	Live Test Q1, Q2, Q3	Sunday 11/12/2017 18:00:01 ET	Sunday 11/12/2017 23:59:59 EST	Tuesday 11/14/2017 23:59:59 EST
Phase 3	Q1, Q2, Q3, Q4	Monday 11/13/2017 00:00:01 ET	Sunday 11/26/2017 15:59:59 EST	-
	Live Test Q1, Q2, Q3, Q4	Sunday 11/26/2017 18:00:01 ET	Sunday 11/26/2017 23:59:59 EST	Tuesday 11/28/2017 23:59:59 EST

Phase 2 Requirements

- Phase 2 accounts for 30% of the total score of the Team Project
 - Phase 1 only accounts for 20%
- You need to continue exploring MySQL and HBase in Phase 2
 - You will continue to work on Q1 & Q2
 - You will work on a new query, Q3
- **You must achieve over 80% correctness, AND at least 50% RPS in BOTH MySQL and HBase in order to get points for each query.**
- Your performance RPS is **SOLELY** determined by the Live-Test
 - Some students cached query results at front-end in phase 1
 - If not done wisely, this may lead to the front-end crashing during the Phase 2 Live-Test
- As before, a report needs to be submitted for Phase 2 after the Live-Test
 - Check the schedule for deadlines

Notes for the Live Test

- During the live test, you must tag your HBase and MySQL cluster with Key: teambackend Value: hbase and Key: teambackend and Value: mysql.
- You must submit both your clusters' DNS before 4 pm, Sunday Nov 12th. Both of your clusters should be ready then.
- You must use the same cluster for all queries. So you can't launch different MySQL clusters for Q2 and Q3.
- Do not launch other testing instances during live test, or else we will count them towards your hourly budget.
- We encourage you to use on-demand instances for the live test or else you run the risk of your instances being shut down unexpectedly
- Leave enough budget for the Live Test. About \$20 should be safe.

Hints for the live test

- The request pattern will differ for Phase 2 and the live test so your solution should handle all types of load.
- Monitor your system efficiently during the live test to recover in case of a system crash.
- Think about what could happen when your cluster should respond to mixed requests of Q1, Q2 and Q3. And try to convince yourself that your design and optimizations are not over aggressive towards some of the queries, and you have reasonable resources (CPU, Disk, Memory, DB&Web Tier) usage.
 - Think and try, but we won't release mixed request before live test

Upcoming Deadlines



- Quiz 10: Unit 5 - Module 18

Due: Thursday, Nov 9th 2017 11:59PM Pittsburgh 

- Team Project - Phase 2

Due: Sunday, Nov 12th 2017 3:59PM Pittsburgh 

- Team Project - Phase 2 - Live Test

Sunday, Nov 12th 2017 4PM-11PM Pittsburgh 

Questions?