15-319 / 15-619 Cloud Computing

Recitation 11

Nov 6th 2018

Overview

- **Last week's reflection**
  - Project 4.1
  - Unit 5 - Module 18
  - Quiz 9
  - Team Project Phase 2 released
- **This week's schedule**
  - Programming exercise on feature engineering
  - Team Project, Phase 2, Queries, 1, 2, 3
  - Team Project, live test
  - Unit 5 - Modules 19 and 20
  - Quiz 10

Reminders

- Monitor AWS expenses regularly and tag all resources
  - Check your bill both on AWS and TPZ
- Piazza Guidelines
  - Tag your question with the correct project
  - Give your submission ID
- Provide clean, modular and well documented code
  - Large penalties for not doing so
- Utilize office hours

P4.1 Reflection

- Programming in Scala on Spark
- Understanding the differences between processing data with MapReduce and Spark
- Exploring Twitter social data with RDD and SparkSQL APIs
- Implementing an iterative processing algorithm - pagerank - on a large dataset
- Utilizing the Spark Web UI to monitor a Spark job and identify performance bottlenecks
- Tuning a Spark program to optimize for time

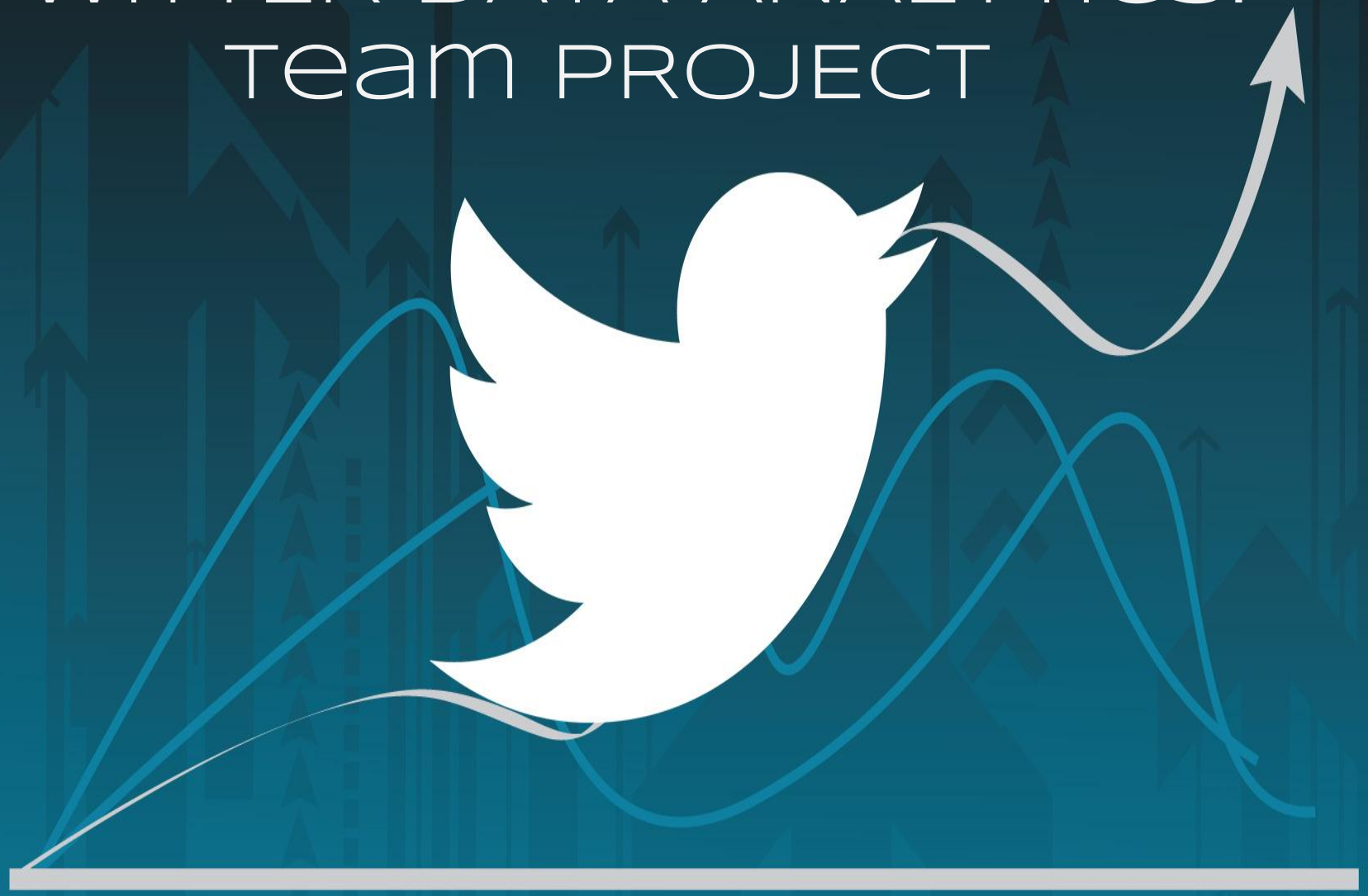
P4.1 Reflection

- Common Issues
  - Handling dangling nodes in the graph
  - Out of memory errors
  - Long running jobs
    - Reduce the amount of data shuffling
- Takeaways
  - Some approaches to implementing pagerank are more efficient than others
  - The Spark Web UI is a useful visualization tool

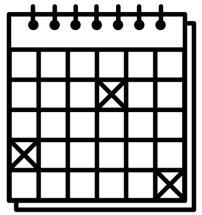
Modules to Read

- UNIT 5: Distributed Programming and Analytics Engines for the Cloud
  - Module 18: Introduction to Distributed Programming for the Cloud
  - Module 19: Distributed Analytics Engines for the Cloud: MapReduce
  - Module 20: Distributed Analytics Engines for the Cloud: Spark
  - Module 21: Distributed Analytics Engines for the Cloud: GraphLab
  - Module 22: Message Queues and Stream Processing

TWITTER DATA ANALYTICS: Team PROJECT



Team Project Time Table



Phase (and query due)	Start	Deadline	Code and Report Due
Phase 1 ● **Q1, Q2**	**Monday 10/08/2018 00:00:00 ET**	Q1: Sunday 10/21/2018 23:59:59 ET Q2: Sunday 10/28/2018 23:59:59 ET	Tuesday 10/30/2018 23:59:59 ET
Phase 2 ● Q1, Q2,Q3	Monday 10/29/2018 00:00:00 ET	Sunday 11/11/2018 15:59:59 ET	
Phase 2 Live Test (**HBase AND MySQL**) ● Q1, Q2, Q3	**Sunday 11/11/2018 16:00:00 ET**	**Sunday 11/11/2018 23:59:59 ET**	**Tuesday 11/13/2018 23:59:59 ET**
Phase 3 ● Q1, Q2, Q3	Monday 11/12/2018 00:00:00 ET	Sunday 12/02/2018 15:59:59 ET	
Phase 3 Live Test (Hbase **OR** MySQL) ● Q1, Q2, Q3	Sunday 12/02/2018 18:00:00 ET	Sunday 12/02/2018 23:59:59 ET	Tuesday 12/04/2018 23:59:59 ET

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)**

Phase 2 Live Test

All times are ET (Pittsburgh).

Remember to submit the DNS! Different DNS for MySQL and HBase....

Submit DNS for Live Test

Information

Time	Task	Description
4:00 pm	HBase	Submit your DNS for the HBase Live Test before the deadline
4:00 pm	MySQL	Submit your DNS for the MySQL Live Test before the deadline
5:30 pm - 5:31 pm	HBase DNS Validation	Validate your HBase DNS. This is the last chance to update your DNS for the HBase Live Test
5:33 pm - 5:34 pm	MySQL DNS Validation	Validate your MySQL DNS. This is the last chance to update your DNS for the MySQL Live Test

Phase 2 Live Test

HBase Live Test

Information

Time	Value	Target	Weight
6:00 pm - 6:25 pm	Warm-up (Q1 only)	0	0%
6:25 pm - 6:50 pm	Q1	28000	6%
6:50 pm - 7:15 pm	Q2	8000	10%
7:15 pm - 7:40 pm	Q3	1500	10%
7:40 pm - 8:05 pm	Mixed Reads(Q1,Q2,Q3)	9000/2500/500	4+5+5 = 14%

Half-time Break

Information

Time	Value
8:05 pm - 8:30 pm	Time to relax and prepare for the MySQL Live Test

Phase 2 Live Test

To get the score for a query (Q2, Q3)

- Correctness >= **80%** for **BOTH** MySQL and HBase
- Achieve **at least 50%** of the target RPS on **BOTH** MySQL and HBase.

MySQL Live Test

Information

Time	Value	Target	Weight
8:30 pm - 8:55 pm	Warm-up (Q1 only)	0	0%
8:55 pm - 9:20 pm	Q1	28000	6%
9:20 pm - 9:45 pm	Q2	8000	10%
9:45 pm - 10:10 pm	Q3	1500	10%
10:10 pm - 10:35 pm	Mixed Reads(Q1,Q2,Q3)	9000/2500/500	4+5+5 = 14%

Team Project Phase 1 Review

- Q1 Task:
  - Front end (web tier) development.
  - QR code encoding and decoding algorithm

- Q2 Tasks:
  - ETL + web tier + database tier,
  - **SELECT** query on MySQL (Relational DBMS),
  - **GET** query on HBase (NoSQL)

Phase 1, Query 1 Tips

- Choose the web framework wisely based on quantitative testing as this will help you in future queries.
- Set an appropriate number of threads to handle the incoming requests.
- Warm up the load balancer adequately before you begin testing.
- Lastly, optimize your code as much as possible.

Phase 1, Query 2 Reflection

- Encoding
- Confusion about how to obtain the description and screen name --- retrieve the latest one from all the contact tweets of the user, ignoring the empty ones.
- For more information, please refer to the good clarification post from piazza

  https://piazza.com/class/jkvtywetsu35vh?cid=2147

Phase 1, Query 2 Tips

- Encoding in MySQL
  - Remember the questions in the checkpoint report?
- Parameter tuning for optimization
  - Buffer pool size
  - Connection pool
  - Query cache
  - setInstances in vertx
  - Number of threads in undertow
- Schema Design
  - Join could be a costly operation

Phase 2, 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, 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 tweets (after censoring)

Phase 2, Query 3 FAQs

Question 1: How to calculate the topic score?

For word ***w*** in the given range of tweets, calculate:

- Calculate the Term Frequency of word *w* in tweet $t^{(i)}$
- Calculate Inverse Document Frequency for word *w*
- Calculate Impact Score of each tweet
- Topic Score for word w =

$$\sum_{i}^{n} TF(w, t^{(i)}) \cdot IDF(w) \cdot ln(Impact(t^{(i)}) + 1),$$

for n tweets in time and uid 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.

Phase 2, Query 3 Tips

- Partition
- Avoiding potential hotspots
- Be aware of the index type for a range search

Reminder

- Your team has a total AWS budget of **$50** for Phase 2
- Your web service should cost ≤ **$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
- Phase 2 - Live Test Targets:
  - Query 1 - 28000 rps
  - Query 2 - 8000 rps (for both MySQL and HBase)
  - Query 3 - 1500 rps (for both MySQL and HBase)

Hints for the live test

- The request pattern will differ for Phase 2 submission test and the live test so your solution should handle all types of load.
- Monitor your system during the live test to recover in case of a system crash.
- Your Phase 2 budget should take into account the cost for the live test.

Upcoming Deadlines

- Feature Engineering Programming Exercise on Cloud9
  - **This week**
- Quiz 10
  - **Due: 11/09/2018 11:59 PM Pittsburgh**
- Team Project : Phase 2
  - **Live-test due: 11/11/2018 3:59 PM Pittsburgh**
  - **Code and report due: 11/13/2018 11:59 PM Pittsburgh**