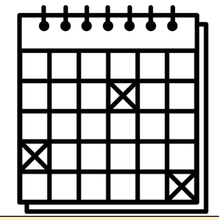
TEAM PROJECT
Twitter Data Analytics

Team Project Time Table



Phase (and query due)	Start	Deadlines	Code and Report Due
Phase 2 • Q1, Q2,Q3	**Monday 03/29/2021 00:00:00 ET**	**Q3 Early Bird Bonus: Sunday 04/04/2021 23:59:59 ET** **Phase2 Due: Sunday 04/18/2021 15:59:59 ET**	
Phase 2 Live Test (Hbase **AND** MySQL) • Q1, Q2, Q3	**Sunday 04/18/2021 16:00:00 ET**	**Sunday 04/18/2021 23:59:59 ET**	**Tuesday 04/20/2021 23:59:59 ET (upload PDF report and verify your submission)**
Phase 3 • Q1, Q2, Q3 (Managed services)	Monday 04/19/2021 00:00:00 ET	Sunday 05/02/2021 15:59:59 ET	
Phase 3 Live Test • Q1, Q2, Q3 (Managed services)	Sunday 05/02/2021 17:00:00 ET	Sunday 05/02/2021 23:59:59 ET	Tuesday 05/04/2021 23:59:59 ET

Team Project Deadlines

- Phase 2 milestones:
  - Q3 Bonus (Reach Q3 target, MySQL+HBase):
    - due on **Sunday, April 4**
  - Phase 2, Live test:
    - Q1/Q2/Q3/mixed on **Sunday, April 18**
  - Phase 2, code, scripts and report:
    - due on **Tuesday, April 20**

Phase 2 Scoreboard (S21)

Submitter	Score	Q1 Score (10)	Q1 Effective Throughput	Q2 Effective Throughput	Q2 Score (50)	Q3 Effective Throughput	Q3 Score (50)
MacUserNoXX	56	10.00	73160.81	6486.71	0.00	3687.94	46.21
CloudCrew	43	10.00	59816.43	15053.60	0.00	4912.45	33.38
Cumulonimbus	31	0.00	0.00	0.00	0.00	1069.91	31.14
HappyCloudHappySurvive	10	10.00	39535.41	9239.71	0.00	4796.25	0.00
CMUeatsthecode	10	10.00	39119.85	0.00	0.00	3119.61	0.00
TheQuinjet	10	10.00	41505.42	0.00	0.00	2287.99	0.00
OnePiece	10	10.00	67118.15	6776.47	0.00	0.00	0.00
Hello	10	10.00	93430.44	0.00	0.00	0.00	0.00
DashDuck	10	10.00	72589.64	0.00	0.00	0.00	0.00
AyoJin	10	10.00	55677.77	0.00	0.00	0.00	0.00
AlphaCC	10	10.00	52847.45	0.00	0.00	0.00	0.00
s21survive	10	10.00	46514.76	0.00	0.00	0.00	0.00
ThreesACloud	10	10.00	43872.92	0.00	0.00	0.00	0.00
GoTeam	10	10.00	42809.37	0.00	0.00	0.00	0.00
HotTamales	10	10.00	40860.81	0.00	0.00	0.00	0.00
Interstellar	10	10.00	36530.94	0.00	0.00	0.00	0.00
GreatPineapple	10	10.00	34197.71	0.00	0.00	0.00	0.00
AceCC	0	0.00	0.00	17912.05	0.00	5145.31	0.00
CCXD666	0	0.00	0.00	7299.93	0.00	1787.29	0.00
cloudcompewpew	0	0.00	0.00	0.00	0.00	2578.96	0.00
Chong	0	0.00	0.00	15851.13	0.00	0.00	0.00
threetimezones	0	0.00	0.00	12212.68	0.00	0.00	0.00
WorstNightmare	0	0.00	0.00	0.00	0.00	728.08	0.00
CudaOutOfMemoey	0	0.00	0.00	2207.26	0.00	0.00	0.00

Query 3 - Tweets Analysis

- **Use Case**: Query 3 analyzes trending topic words within a specific range of timestamps and users. It also ranks and returns the most trending tweets.

- **Query**:
  GET
  /q3?uid_start=<left_bound_uid>&uid_end=<right_bound_uid>
  &time_start=<left_boudn_uid>&time_end=<right_bound_uid>&n1=<max_topic_words>&n2=<max_tweets>

- **Response:**
  <TEAMNAME>,<AWSID>\n
  $word_1$:$score_1$\t$word_2$:$score_2$\t…\t$word_{n1}$:$score_{n1}$\n
  $impactScore_1$\t$tid_1$\t$text_1$\n
  $impactScore_2$\t$tid_2$\t$text_2$\n
  …
  $impactScore_{n2}$\t$tid_{n2}$\t$text_{n2}$\n

Target Throughput: 1,500 RPS for both MySQL and HBase

Query 3 - What words are trending?

- For one given word, one tweet will have a measurement of the word's importance
  - named *TF-IDF Score*: its importance within the tweet **multiplies** its importance among all the queried tweets

- Then we multiply this TF-IDF score with that tweet's importance
  - named *Impact Score*: EWC * (favorite_count + retweet_count + followers_count)

- The final popularity of the word will the sum of above productions among all the queried tweets
  - sum(Xi * ln(Yi + 1)) where i from 1 to n

- Sort the words by the score and return the first n1 ones

Query 3 - What tweets are trending?

- Easy, use the *Impact Score* defined previously
  - EWC * (favorite_count + retweet_count + followers_count)

- Sort the tweets by the score and return the first n2 ones

Query 3 - Filtering

- Similar to query 2, please remove duplicate tweets and malformed tweets; however, in query 3, you don’t filter the tweets without hashtags

- The language allowed in query 3 is **en**.

Query 3 - What is a *word*?

- URL
  - Remove short URLs when calculating impact scores and topic words
  - Keep the short URLs when returning the popular tweets
  - Use the regular expression we provide

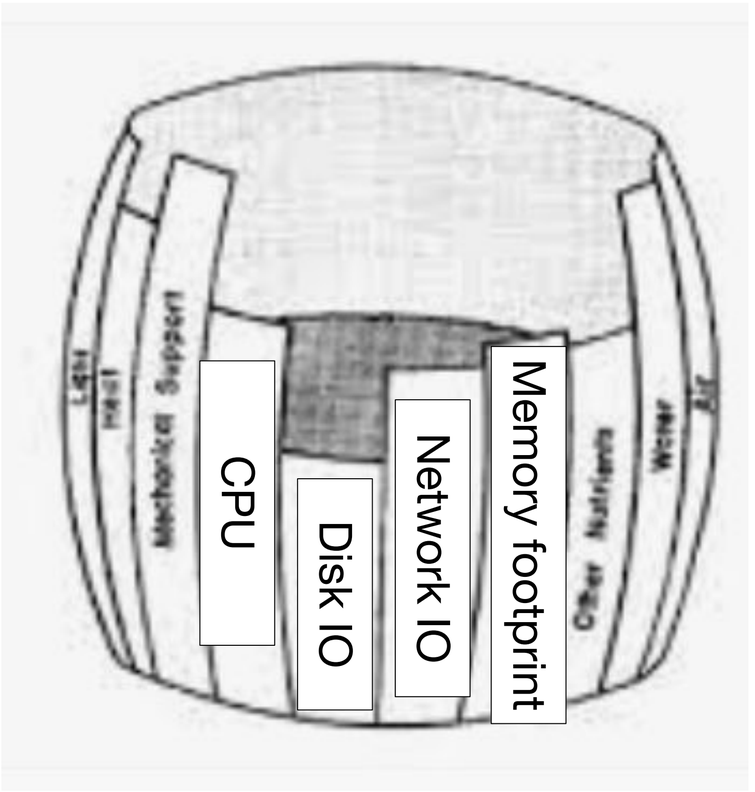
- Text Censoring - is bad word also a word?
  - Keep the bad words when calculating impact scores and topic words
  - Censor the bad words when returning the topic words and popular tweets
  - Note: the way we define a bad word is different from the definition of words when calculating impact scores and topic word scores.

Query 3 - Hints

- [correctness] Implement ETL on the first part of the dataset and compare to the provided reference files
- [correctness] Same as query 2, please test on the mini dataset and verify your results to the reference mini server.
- [budget] Use Azure/GCP for ETL as much as possible since you have limited budget on AWS
- [design] Read the report before you start! It can give you a direction for your development

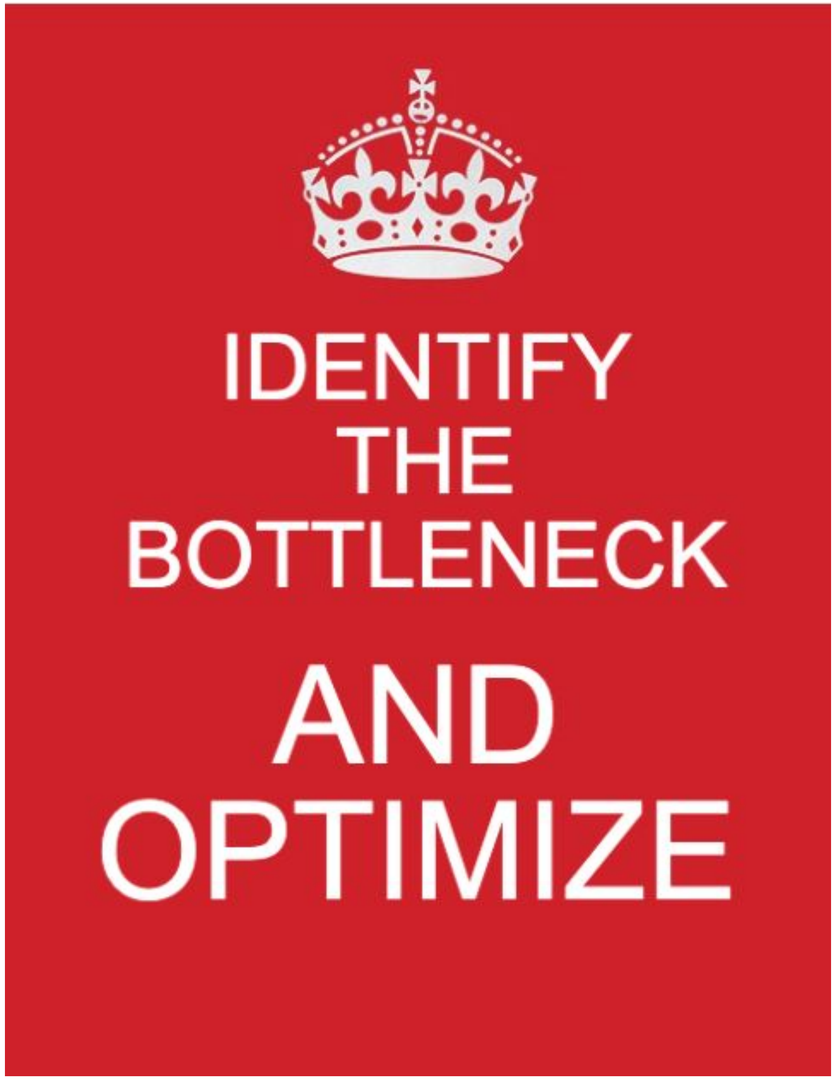
From functioning to high-performance

- No profiling, no optimization
- Test, instead of guess
- If your RPS is less than the half of target RPS...
  - Write a breakdown of your query process
    - Which part takes the most of time?
    - Can you halve the processing time?
  - Also, improve your correctness in parallel



Hourly Budget Reminders

- **$0.73/hour (MySQL) and $0.89/hour (HBase)** apply to all submissions and the live test
- From the hourly budget, it includes EC2(CPU), EBS(disk storage), ELB(load balancing).
- EMR pricing is an additional service fee for AWS managing the Hadoop cluster. It's **excluded** in the hourly budget, but 1) it doesn't mean you get CPU and disk storage for free, 2) you still pay this fee from your team AWS account
- Calculate your EBS carefully to avoid exceed the hourly budget
- Total budget on AWS is $80 for the whole Phase 2, including the live test



Have fun!