Recitation - 11/02

Team Project Phase 2

Tom Chiu

Agenda for Phase 2

- Phase 2 lasts for 2 weeks. (11/02 - 11/15)
- Live Test for Phase 2 takes place on 11/15.

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

Please read the writeup for more details of the live test schedule

Query 3 - Tweets Analysis

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

- **Use Case**: Query 3 allows users to analyze trending topic words within a range of timestamps and users. Also, it ranks the tweets with the impact scores

- **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

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 - Short URL

- When you **calculate** impact scores and topic words, please remove short URLs.

- But when you return the tweets, please DO NOT remove them, i.e. keep the short URLs in the original tweet.

  - python: `(https?|ftp):\/\/[^\t\r\n /$.?#][^\t\r\n ]*`
  - java: `(https?|ftp)://[^\\t\\r\\n /$.?#][^\\t\\r\\n ]*`

Query 3 - Text Censoring

```
15619ppgrfg
4e5r
5u1g
5uvg
a1ttn
a1ttre
abournq
abowbpxl
abowbxrl
ahgfnpx
ahzoahgf
```

- “Bad” words occur in tweets, but we don’t want to expose them.
- Censor the original tweet before short URL elimination.
- The list of banned words are encrypted by ROT13. Please decrypt the words by yourselves, e.g. **15619ppgrfg** will be decrypted as **15619cctest**
- Rule for censoring: words are separated by non-alphanumeric characters, including ‘ and **-**. *NOTE: The definition of words here is different from impact scores and topic words calculation.*

```
  I love Cloud compz... cloud TAs are the best... Yinz shld tell yr frnz: TAKE CLOUD COMPUTING NEXT SEMESTE
R!!! Awesome. It's cloudy tonight.
```

```
  I love C***d compz... c***d TAs are the best... Yinz shld tell yr frnz: TAKE C***D COMPUTING NEXT SEMESTE
R!!! Awesome. It's cloudy tonight.
```

Query 3 - Word Definition for Score Calculation

- For impact scores and topic words calculation, a word is defined as followed
  1. Containing one or more consecutive **alphanumeric** characters
  2. With zero or more ‘ or/and **-** characters
  3. But must have at least one alphabet

  So words are separated by all other characters not mentioned above.

- Can you count how many words in the following tweets?

```
Query 3 is su-per-b! I'mmmm lovin' it!
```

Query 3 - Word Definition for Score Calculation

- For impact score and topic words calculation, a word is defined as followed
  1. Containing one or more consecutive **alphanumeric** characters
  2. With zero or more ‘ or/and **-** characters
  3. Not containing only numbers and/or ‘ and/or **-**

  So words are separated by all other characters not mentioned above.

- Can you count how many words in the following tweets?

```
Query 3 is su-per-b! I'mmmm lovin' it!
```

Query 3 - Impact Score Calculation

```
a
about
above
across
after
afterwards
again
against
all
```

- **Precondition**: Please remove the short URLs before the calculation.
- Same as banned words, stop words are case-insensitive.
- Effective Word Count is the total number of non-stopwords.

```
I love Cloud Computing
```

- A tweet has **favorite_count** and **retweet_count** fields.
  A user has a **followers_count** field.
  If a field does not exist, the field is set to 0.

```
impact_score = EWC * (favorite_count + retweet_count + followers_count)
```

- For negative impact scores, please set them to 0

Query 3 - Impact Score Calculation

```
a
about
above
across
after
afterwards
again
against
all
```

- **Precondition**: Please remove the short URLs before the calculation.
- Same as banned words, stop words are case-insensitive.
- Effective Word Count is the total number of non-stopwords.

```
I love Cloud Computing
```

EWC = 3!

- A tweet has **favorite_count** and **retweet_count** fields.
  A user has a **followers_count** field.
  If a field does not exist, the field is set to 0.

```
impact_score = EWC * (favorite_count + retweet_count + followers_count)
```

- For negative impact scores, please set them to 0

Query 3 - Topic Words Extraction

- Precondition: Please remove the short urls before the calculation.
- Text censoring should be done **AFTER** topic words extraction `cloud` `could`
- We use a variant of TF-IDF for topic word calculation
  TF: term frequency word **w** in a tweet **t** (counts of word **w** / counts of total words)
  IDF: ln (*Total number of tweets in range / number of tweets with **w** in it*)
- Aggregate all tweets in range to get the topic score for a word **w**
  where **Xi** is the TF-IDF score of word **w** in Tweet **Ti**
  and **Yi** is the calculated impact score of **Ti**

```
sum(Xi * ln(Yi + 1)) where i from 1 to n
```

- **Note:** stopwords are not considered as topic words. But they're counted as words in calculation of total number of words in a tweet.

http://www.tfidf.com/

Query 3 - Hints for ETL

- Implement ETL on the first part of the dataset and compare to the provided reference files
  Note: The reference file is for verifying filtering and the format of timestamps. It does not give any suggestion for ETL or database designs

- Please finish project 4.1 as soon as possible. You will learn Azure Databricks in the project, and it is highly recommended to try out **Azure Databricks** for Apache Spark

- Same as query 2, please test on the mini dataset and verify your results to the reference mini server.

- Still, please use Azure/GCP for ETL as much as possible since you have limited budget on AWS

Query 3 - Other hints and reminders

- Try to diagnose the bottleneck and improve it. Profile your server to have more ideas!
- Read the report before you start! It can give you a direction for your development
- Only instances of **M** family are allowed, and they cannot be larger than **large** type
- Calculate your EBS carefully to avoid exceed the hourly budget
- **$0.73/hour (MySQL) and $0.89/hour (HBase)** apply to all submissions and the live test
- Total budget on AWS is $60 for the whole Phase 2 including the live test

Recommended Timeline for Phase 2

Phase 2 Weeks	Tasks	Deadline
Week 1 (11/02 - 11/08)	• **Fix Q1 and Q2 if you did not achieve basic target** • **Complete Q3 ETL & Initial Schema Design** • **Achieve Q3 basic correctness and make submissions on TPZ for both MySQL and HBase**	**Early Bird Bonus for Query 3: 11/08 No hand-in report, but please start early to leave room to reiterate your designs**
Week 2 (11/09 - 11/15)	• Achieve close to 100% correctness for both Q3 MySQL and HBase • Optimize MySQL and HBase servers to achieve target throughput	Final Report due on 11/17
Live Test (11/15 4 pm - 10:35 pm)	• Set up MySQL and HBase servers before 4 pm (loading data, warming up, etc.) • Monitor server status during the live test	• Submit your DNS endpoint for HBase and MySQL at 4 pm