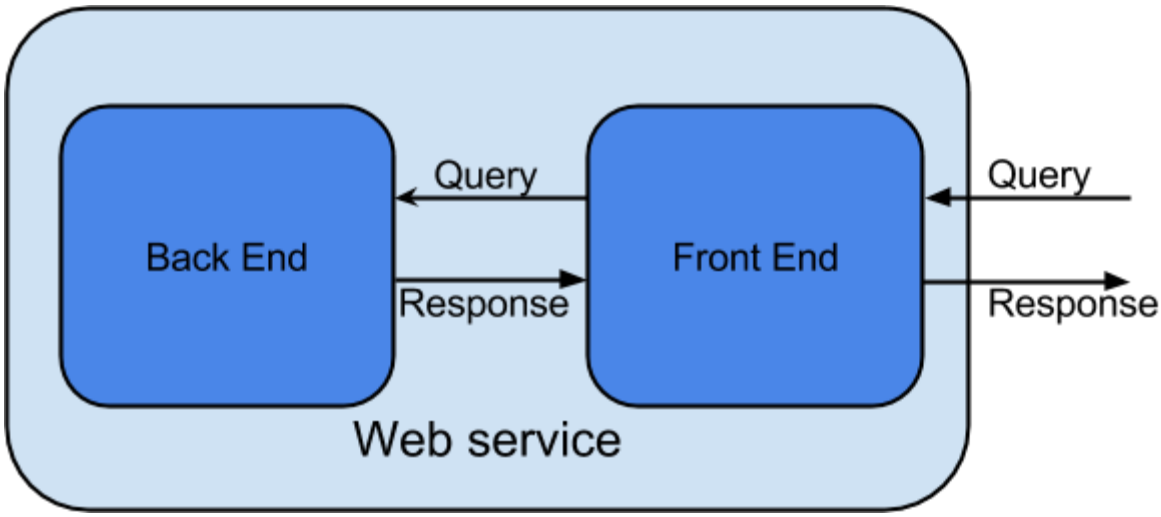
Twitter Analytics Web Service
A 619 Competition

Introduction

After making a huge profit for the Massive Surveillance Bureau, you realize that your Cloud Computing skills are being undervalued. To maximize your own profit, you launch a cloud-based startup company with one or two colleagues from 15619.

A client has approached your company and several other companies to compete on a project to build a web service for Twitter data analysis. The client has provided hundreds of gigabytes of raw tweets for your consumption. Your responsibility would be to design, develop and deploy a web service that meets the throughput, budget and query requirements of the client.

Your team needs to build (and optimize) two components:



1. **A Front end:** This should be a web service able to receive and respond to queries. Specifically, the service should handle incoming HTTP requests and provide suitable responses (as defined in the **Query Types** section below). The interface of your service is REST-based.
    a. Users access your service using an HTTP GET request through an endpoint URL. Different URLs are needed for each query type, which are shown in **Query Types** section. Query parameters are included within the URL string of the HTTP request.
    b. An appropriate response should be formulated for each type of query. The response format should be followed exactly otherwise your web service will not provide acceptable responses when tested with our load generator and testing system.
    c. The web service should run smoothly for the entire test period, which lasts for several hours.
    d. The web service must not refuse queries and should tolerate a heavy load.

2. **A Back End:** This is used to store the data to be queried.
    a. You will evaluate SQL (MySQL) and NoSQL (HBase) databases in the first two phases of this project. You should compare their performance for different query types for different dataset sizes. You can then decide on an appropriate storage back end for your final system to compete against other systems.

3. Your web service should meet the requirements for throughput and cost for queries at a provided workload.

4. The overall service (development and deployment test period) should cost under a specified budget. Less is generally better, otherwise your competitor will win the contract.

5. Your client has a limited budget. So, you can **ONLY** use **m1.small**, **m1.medium**, **m1.large, m3.medium and m3.large** instances for your web service (both your front-end and back-end systems). However, you can use spot instances for batch jobs and development but not for the deployed web service (during the live test period).

Objectives:
The goal of this project is to integrate everything that you have learned from the course (and many, many new things) in designing, developing and deploying a real working cloud-based web solution. We encourage you to discover and utilize whatever tools you find. If the tools do not appear in the design constraints of this handout, you **MUST** discuss them with the professor or TAs before using them.

Dataset
The dataset collected by the client is available on **S3**. For phase 1, you should use the folder:

Input and output
The web service solution should provide responses to specific queries on the twitter dataset. Users can submit queries about tweets based on userids, tweet id, tweet time, location, hashtags and photos. The client has collected billions of tweets through Twitter's streaming API in JSON formatted files, which are stored on S3.

The input is in a JSON format. Each line is a JSON object representing a tweet. Twitter's documentation has a good description of the data format for tweets and related entities in this format. https://dev.twitter.com/docs/platform-objects/tweets.

Timeline

This project has 3 phases. In each phase you are required to implement your web service to provide responses to some particular types of queries. The query types will be described in the writeup of each phase.

At the end of each project phase, you need to submit some deliverables including:

1. Performance data of your web service.
2. Cost analysis of the phase
3. All the code related to the phase
4. Answer to questions associated with the the phase.
5. A report for each phase.

The report must include detailed reasoning for your design decisions and deliverables for each phase.

The time allotted for the phases is as follows:

Phases	**Duration**	**Query Type**
Phase 1	2 weeks	Q1, Q2
Phase 2	2 weeks	Q1, Q2, Q3, Q4
Phase 2 Live Test	6 hours	Q1, Q2, Q3, Q4, mix-Q1Q2Q3Q4
Phase 3	2 weeks	Q1, Q2, Q3, Q4, Q5, Q6
Phase 3 Live Test	6 hours	Q1, Q2, Q3, Q4, Q5, Q6, MIX-Q1Q2Q3Q4Q5Q6

Web Service Load Generation and Testing System

In this project, your web service interacts with the 15619 project testing system developed by us to test and grade your service.

The testing system has a front end at: http://15619project.org

Registration

You should finish team registration by **Saturday of week 0 (Oct 4th, 2014)** . You need to provide a valid andrew id in order to confirm your registration. You will receive an email to log in and set your password in the first week.

Scoreboard

We provide a real-time scoreboard for you to examine how your performance compares to other teams. Your **best submission** for each query will be displayed on the scoreboard.

Submit your request

You can submit a test request at any time. Please use the **public DNS address** of your instance or ELB so that the testing instance can reach your service. We provide different testing period lengths, please think carefully about what you need to verify for each request and choose an appropriate duration.

- How is my request tested?
  When you submit a request, a testing instance will be launched by us. This instance will run a benchmark program for the specified query to the provided address in order to test the performance and correctness of the web service. After the testing period ends the results will be displayed in your submission history table.

Wait List

Since we have a limited amount of resources, if there are many teams submitting requests at the same time, your request may be waitlisted. You can check to see how long you need to wait before your request starts running. Our system will try to add or deduct resources automatically based on the current average waiting time, so don't worry if you see a long waiting time.

- Why can't I submit a request?
  To save cost and prevent repeat submissions, every team can only have **one request running or waiting**. In other words, if anyone in your team submits a request, no one in your team can submit another request until the running job finishes.

- Can I cancel my request if I modify my service and want to re-submit?
  Yes, you can.

Submission History

When your request is running, you can check the submission history page to see the progress of the current request and how many seconds are remaining. After your request finishes, you your results will be displayed. You can also view all your previous submissions. If you have doubts for a specific query, you can take down the submission ID and contact us for more details.

- How to interpret my result?

There are several numbers for each result:

Throughput: average queries per seconds during the testing period

Latency: Average latency for each query issued during the testing period

Error: The error rate. Your message type in HTTP header of your response must be 2xx. Otherwise it will be recognized as an erroneous response and counted in error rate.

Correctness: This column indicates whether your service responds with the correct result. Your response should exactly match our standard response so please read the format requirement of each query carefully.

Score: This is your final score of this request. You score is based on effective throughput. The effective throughput is calculated as follows:

- **Effective Throughput = Throughput * (100 - Error / 100) * (Correctness / 100)**
    - As you can see, error and correctness can severely impact your effective throughput

Live Test

Phases 2 and 3 culminate with a Live Test, where all systems are simultaneously tested in a single window. Your grade for both phases is based on your performance in these tests.

Score Calculation

- The score is a weighted sum of your raw score for each query
- Raw Score = Effective Throughput/Target Throughput
    - The Target Throughput will be provided by us

Bug Report

If you encounter any bugs in this sytem or have any suggestions for improvement, feel free to use this form to help us improve.

Phase 1

Introduction

In the first phase of the 619 project, you will build two web services from scratch. Each web service has to respond to two types of queries, with data fetched from a storage system, which you must design and control. Each web service connects to a different storage system. A query is an input generated by a test system, which requires a fixed response. Grading depends on the accuracy and performance of your web service's response to these queries.

This phase is designed to ensure that you have the required skills to cope with more complicated queries, so please allocate enough time to finish the requirements by the deadline.

Specifically, you will design and develop a front-end system (a highly parallel, concurrent web server) that can attach to either of two back-end systems (HBase and MySQL). In this phase you will be expected to learn about the advantages and disadvantages of each type of back-end database system.

A report must be written for this phase that conforms to a template.

Dataset

The dataset for this phase is a folder containing about 600 GB of JSON-format tweets. For more detail explanation please refer to the overall project writeup or the Twitter API.

The dataset is at s3://15619f14twittertest2/600GB

Phase 1 Query types:

Your web service's front end will have to handle the following query types through HTTP GET requests **on port 80 [you cannot use any other port]**:

- ## Heartbeat and Authentication (q1)

  The query asks about the state of the web service. The front end server responds with the project team id, AWS account ids and the current timestamp. It is generally used as a heartbeat mechanism, but it could be abused here to test whether your front end system can handle varying loads.

  For each request to q1, the load generator will send a large numeric key. You need to use this as follows to generate the response:

  a. We give you a public "key" (which is the large prime number below):

  - X=687676683235176539649637753447605000297085748381526291845035586985008516705339467263431539122405215 3

b. The load generator creates a number X*Y, where Y is another (possibly) huge prime number. X*Y is sent to your web service.
c. You need to divide this number (X*Y) by X to retrieve Y
d. You then authenticate by sending Y back to the server as a part of your response

- REST format: GET /q1?key=<large_number XY>
  - Example:

    http://webservice_public_dns/q1?key=20630300497055296189489132603428150

    00891257245144578875535106760955025550116018401790294617367215645 9

- response format:
  <THE NUMBER Y>
  TEAMID,AWS_ACCOUNT_ID1,AWS_ACCOUNT_ID2, AWS_ACCOUNT_ID3
  yyyy-MM-dd HH:mm:ss

- Example (for the URL above):
  3
  TeamCoolCloud,1234-0000-0001,1234-0000-0002,1234-0000-0003
  2004-08-15 16:23:42

  [There is a \n after the last line]

- Minimum throughput requirement: 15000 qps

The following public key is fixed for this project:
X=68767668323517653964963775344760500029708574838152629184503558698500851670533946726343 15391224052153

- **Text Cleaning and Analysis (q2)**
  This query asks for the tweet(s) posted by a given user at a specific time. This query tests that your backend database is functioning properly.

  We send you a user id and tweet time, and you need to send us three pieces of data about the relevant tweet(s).
  a. The tweet ID
  b. The censored text
  c. The sentiment score of that tweet

  Each of these are explained in the following paragraphs.

  (Link to Twitter API)

Step 1: The tweet id can be fetched from the 'id' or the 'id_str' field.
Step 2: The tweet text can be fetched with the 'text' field. There are two processes that must run on each tweet text. To do them, you need to split the string into separate "words" wherever a non-alphanumeric character is encountered.

A) Sentiment Score Calculation:
We provide a list of lowercase English words that have a corresponding "sentiment" score. For each word in the text, convert it to lowercase and check if it has a sentiment score. If it exists, add the corresponding sentiment value to the score for that tweet. Scoring for all tweets starts with a zero.
This list is from the AFINN dataset.[1]
AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. The file is tab-separated.
DOWNLOAD IT HERE.

B) Text Censoring
Since we have innocent little undergrads also doing this assignment, we did not want to expose them to some of the language that exists on Twitter. So we provide a mechanism to sanitize the text. We give you an ROT13ed version of all banned words that should not occur in your output text. [ROT13 is a simple letter-based substitution cipher. **For instance, the first line in the list of banned words is 15619ppgrfg, which means that the first banned "word" is 15619cctest**.
This list is filtered from last year's students' colorful posts on Piazza near the end of the 15619 (all these students started working on their project late).[2]
DOWNLOAD IT HERE.

Ensure that in both cases, you use our files (links above) and do not download the original, as we have done some cleanup to make it easy to use for your code.

You must do the steps in order, that is, first calculate the sentiment for a word and then (if required) censor it. A word is censored by replacing the inner characters by asterisks (*).

[1] Hansen, L. K., Arvidsson, A., Nielsen, F. Å., Colleoni, E., & Etter, M. (2011). Good friends, bad news-affect and virality in twitter. In *Future information technology* (pp. 34-43). Springer Berlin Heidelberg.
[2] No, actually it's mostly from http://www.bannedwordlist.com/. Decrypting and reading this file could be entertaining and informative as long as you do not use what you learn on the course staff

For e.g. assume that the word **cloud** is on our banned list.
Consider the following tweet:

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

When you reply, it should have the format:

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

And it should have a score of +10 (Best = +3, Love = +3 and Awesome = +4)

Notice the case of all the uncensored letters is maintained. Also notice that "cloudy" is uncensored (since it is not on the list of banned words).

- Please note that since the tweet texts in the JSON objects are encoded with unicode, different libraries might parse the text slightly differently. To ensure your correctness, we recommend that you use the following libraries to parse your data.
    - If you are using Java, please use simple json/gson
    - If you are using Python, use json module in standard library

  You can choose to do all these calculations in the ETL (remember that it will drive up the cost and duration of the MapReduce job) or at the end when the query is requested (if you can write fast code).

- REST format: GET /q2?userid=uid&tweet_time=timestamp
    - Example: http://webservice_public_dns/q2?userid=123456789&tweet_time=2004-08-15+16:23:42
- response format:

```
TEAMID,AWS_ACCOUNT_ID1,AWS_ACCOUNT_ID2,AWS_ACCOUNT_ID3
Tweet ID1:Score1:TWEETTEXT1
Tweet ID2:Score2:TWEETTEXT2
...
```

(Sort numerically by Tweet ID. There should be a line break '\n' at the end of each response.)

- Example:

```
WeCloud,1234-5678-9012,1234-5678-9013,1234-5678-9014
12345678987654321:2:When I met Mr Mandela it was one of the most memorable days of my life. A truly inspirational human being....
1234567898765432 3:0:....He will live on in my heart forever. R.I.P
```

[There is a \n after the last line]

- Minimum throughput requirement: 7000 rps

Tasks

- *Make sure you tag all the instances you use in this phase with the key "**15619project**" and value "**phase1**". Remember the tags are **case-sensitive**.* Missing and incorrect tags will lead to unnecessary penalties. Just make it a habit to tag correctly.

Task 1: Front End

This task requires you to build the front end system of the web service. The front end should accept RESTful requests and send back responses. For phase 1, the front end system is only required to handle q1 and q2. q1 does not require connectivity to a database, while q2 needs HBase and MySQL (both, separately).

Design constraints:

- Execution model: you can use any web framework you want.
- Spot instances are highly recommended during development period, otherwise it is very likely that you will exceed the overall budget for this step and fail the project.

Recommendation:

You might want to consider using auto-scaling because there could be fluctuations in the load over the test period. To test your front end system, use the Heartbeat query (q1). It will be wise to ensure that your system comes close to satisfying the minimum throughput requirement of heartbeat requests before you move forward. However, as you design the front end, make sure to account for the cost. Write an automatic script, or make a new AMI to configure the whole front end instance. It can help to rebuild the front end quickly, which may happen several times in the building process. Please terminate your instances when not needed. Save time or your cost will increase higher than the budget and lead to failure.

Task 2: ETL

This task requires you to load the Twitter dataset into the database using the **extract**, **transform** and **load** (ETL) process in data warehousing. In the extract step, you will extract data from an outside source. For this phase, the outside source is a JSON Twitter dataset of tweets stored on **S3,** containing about 200 million tweets. The transform step applies a series of functions on the extracted data to realize the data needed in the target database. The transform step relies on the schema design of the target database. The load phase loads the data into the target database.

You will have to carefully design the ETL process using AWS resources. Considerations include the programming model used for the ETL job, the type and number of instances needed to execute the job. Given the above, you should be able to come up with an expected time to complete the job and hence an expected overall cost of the ETL job.

Once this step is completed, you should backup your database to save cost. If you use EMR, you can backup HBase on S3 using the command:

./elastic-mapreduce --jobflow j-EMR_Job_Flow --hbase-backup --backup-dir s3://myawsbucket/backups/j-EMR_Job_Flow

This backup command will run a Hadoop MapReduce job and you can monitor it from both Amazon EMR debug tool or by accessing the Jobtracker. To learn more about Hbase S3 backup, please refer to the following link http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-hbase-backup-restore.html

Design constraints:

- Programming model: you can use any programming model you see fit for this job.
- AWS resources: You should use SPOT instances for this step otherwise it is very likely that you will exceed the budget and fail the project.

Recommendations: For this phase, the ETL jobs will utilize a small dataset (600 GB), however, in later phases, you will be required to run ETL for a much larger dataset. Always test the correctness of your system using a tiny dataset (example 200MBs). If your ETL job fails or produces wrong results, you will be burning through your budget. After the database is populated with data, it will be wise to test the throughput of your back end system for different types of queries. Ensure that your system produces correct query responses for all query types.

Task 3: Back end (database)

For your system to provide responses for q2, you need to store the required data in a back-end database. Your front end system connects to the back end and queries it in order to get the response and send it to the requester.

You are going to use both **HBase** and **MySQL** in this phase. We will provide you with some references that will accelerate your learning process. You are expected to read and learn about these database systems on your own in order to finish this task. You also need to know enough about the advantages and disadvantages of each database system in order to answer the questions in the Phase 1 report template.

This task requires you to build the back end system. It should be able to store whatever data you need to satisfy the query requests. You should use spot instances for the back end system development. But you can use on-demand prices during the live-test periods of your system in future phases of this project.

We require you to build two back ends, one with **HBase** and the other with **MySQL** as your

back end databases. You need to consider the design of the table structure for the database. The design of the table significantly affects the performance of a database.

In this task you should also test and make sure that your front end system connects to your back end database and can get responses for queries.

Design constraints:

- Storage model and justification: you should use both HBase and MySQL and explain the differences between them.
- Spot instances are highly recommended during the development period, otherwise it is very likely that you will exceed this phase's overall budget and fail the project.

Recommendations:
Test the functionality of the database with a few dummy entries before the Twitter data is loaded into it. The test ensures that your database can correctly produce the response to the q2 query.

Budget for Phase 1
You will have a budget **$35/team** for all the tasks in this phase. Again, make sure you tag all of your instances in this phase with key "***15619project***" and value "***phase1***". Additionally, all instances in your HBase cluster should be tagged with: **"15619backend"** and value **"hbase"**. And all instances with MySQL installed should be tagged with: **"15619backend"** and value **"mysql"**. It is up to you to decide how to spend the budget in order to get the best performance.

Each run (test submitted to the website) must have a maximum budget of $1.25 (include on-demand EC2, storage and ELB costs. Ignore EMR and Network, Disk I/O costs).

Recommendations:

1. Use smaller instances when you do functionality development and verification of the front end (you may even finish a lot of the development needed for this task on your own laptop).
2. Think carefully about your database schema to avoid running ETL too many times.
3. Spend some time to develop a plan on how to allocate the budget so that you can complete all tasks and get good performance.
4. This is an opportunity for you to practice working as a team. For example, assign clear tasks to team members, communicate regularly, update each other on progress, specify clear interfaces, etc… All members should pull their own weight.
5. Have fun! This should be a challenging but a rewarding learning experience.

Deliverables

Phase 1

You need to make sure you have **at least 1 test record on the testing system for q1 and 2 for q2 respectively which meet the minimum throughput requirements**. For q2, you need to have at least two submissions, one for MySQL and another for HBase. You just need to make both of them pass the Target Throughput. The submission time of the record on the testing system must be before the deadline of this phase. Besides, please submit

1. Phase 1 checkpoint report in PDF format (Phase1 checkpoint report template)
2. All your code written in this phase (front end, ETL, etc...)

To submit your work:

1. Package your completed Phase 1 checkpoint report and all the source code files you wrote in the phase into a single TEAM_NAME.zip file.
2. Upload the TEAM_NAME.zip file to the testing system at this URL.

Grading

Phase 1 accounts for 10% of the total grades for this 15619Project. You need to finish all the tasks in order to move on to the next phase which will add new queries. Your success on the next phases is highly dependent on how much you achieve in Phase 1. So, even though it only accounts for 10%, Phase 1 has a huge impact on the overall success of the project and should be taken very seriously.

Hints and References

Front End Suggestions:

Although we do not have any constraints on the front end, performance of different web frameworks vary a lot. Choosing a slow web framework may have a negative impact on the throughput of every query. Therefore, we strongly recommend that you **do some investigation about this topic** before you start. Techempower provides a very complete benchmark for mainstream frameworks, you may find it helpful.

You can also compare the performance of different frameworks under our testing environment by testing q1 since it is just a heartbeat message and has no interactions with the back end. Please also think about whether the front end framework you choose has API support for MySQL and HBase.

General Guidelines for ETL:

How to do ETL correctly and efficiently will be a critical part for your success in this project. Notice that ETL on a large dataset could take 10 - 30 hours for just a single run, so it will be very painful to do it more than once, although this may be inevitable since you will be refining your schema throughout your development.

You may find your ETL job extremely time consuming because of the massive dataset and the poor design of your ETL process (we will release an even larger dataset in the later phases). Due to many reasons that lead to failure of your ETL step, please start thinking about your database schema and doing your ETL as EARLY as possible.
Also try to utilize parallelism as much as possible, loading the data with a single process/thread is a waste of time and computing resources, MapReduce may be your friend, though other methods work so long as they can accelerate your ETL process.

Reference documents for MySQL and HBase

MySQL

- http://dev.mysql.com/doc/
- OLI Unit 4 and Project 3

HBase:

- https://hbase.apache.org/
- OLI Unit 4, Module 14
- OLI Project 3 and Project 4