15619 Project Phase 1 Report

Performance Data and Configurations

	Front end	Web service with HBase	Web service with MySQL
Query	q1	q2 (small dataset)	q2 (small dataset)
Scoreboard request ID	7946	10150	10569
Instance type	Medium	Large	Large
Number of instances	8	8	1
Queries Per Second (QPS)	12004.3	822.4	401.6
Error rate	0	0	0.01%
Correctness	100	100	76%
Cost per hour	0.56	1.12	0.140

Task 1: Front end

Questions

1. Which front end system solution did you use? Explain why did you decide to use this solution.
We used Java Servlet with Tomcat. We also tried PHP, which turns out to be really slow when processing the big integer. While Java Servlet with Tomcat can handle this problem and it is easy to implement.

2. Explain your choice of instance type and numbers for your front end system.
A medium or large instance.
(1)For q1, medium instances is enough to handle requests. We used ELB to link 6 -10instances together.
(2)For q2 MySQL, We tried m3.large instances to get faster database query.
(3)For q2 HBase, We built a large instance which is linked to the master instance of EMR. Because it is the important part to receive and send a large number of information.

3. Did you do any special configurations on your front end system? Explain your design decisions and provide details here.
We have no special configuration on front end system. But we put some of the string processing in the front end. This is because our design of the ETL was not good enough and we didn’t have

time to change the format. We doubt this may undermine our performance greatly.

4. What is the cost to develop the front end system.
We usually develop the instances using micro type. And we are stilling within the limit of free tier, so there is no cost for front end.

Task 2: Back end (database)

Questions
1. Describe your table design for both HBase and MySQL. Explain your design decision.
For MySQL, we have the following table design

```
+---------+-------------+------+-----+---------+-------+
| Field   | Type        | Null | Key | Default | Extra |
+---------+-------------+------+-----+---------+-------+
| userid  | bigint(20)  | NO   | MUL | NULL    |       |
| time    | varchar(40) | NO   | MUL | NULL    |       |
| tweetid | bigint(20)  | YES  |     | NULL    |       |
| score   | int(11)     | YES  |     | NULL    |       |
| content | text        | YES  |     | NULL    |       |
+---------+-------------+------+-----+---------+-------+
```

We created index for userid and time so we can quickly do the database query.

In Hbase, the row key is userID with time, the value is the tweetID with score and text. This pair of row key and value are stored in table.(value are a column family with 3 columns)

```
+-----------+-------------------------------------------------+
| row key   | tweet_time+userid (2013-09-12+12:34:231234567) |
+-----------+-------------------------------------------------+
| qualifier | tweet_id                                        |
+-----------+-------------------------------------------------+
| value     | score:text                                      |
+-----------+-------------------------------------------------+
```

2. What is the cost to develop your back end system.
For MySQL, one m3.large instance running for ~ 20 hours used for implementing, testing and debugging. ~$ 2.8
For Hbase, we have built an EMR with one m1.large master and 2 m1.large slaves to develop. They ran for 6 hours, which cost $3.942. And then I used the emr to load data into Hbase with ten slaves and one master running for 1 hour, which cost $2.409. So it cost $6.351

Task 3: ETL
Since ETL was performed for both HBase and MySQL, you will be required to submit information for each type of database.

MySQL:

1. The code for the ETL job
   Please refer to the attached phase1_mapper.py, phase1_reducer.py and phase1_sql.sql

2. The programming model used for the ETL job and justification
   We used python programs doing the mapreduce work to extract and transform the data into .csv files containing only the information we needed for this job. This job dramatically reduced the amount of data to be stored in the database. (from ~600G to ~ 27.3G)

3. The type of instances used and justification
   We used ~14 m1.large instances for the mapreduce streaming program. It takes us ~6.5 hours in total. (we increased the number of instaces to 20 in the last two hours). I think the amount of time it take is a little too long so we tried to use the largest instance allowed in this phase.

4. The number of instances used and justification
   Please refer to question 3

5. The spot cost for all instances used
   m3.medium 0.02

6. The execution time for the entire ETL process
   EMR: ~6.5 hour
   Data loading: ~ 45 min.
   Building index: ~1 hour
   Tota: ~ 8.25 hour

7. The overall cost of the ETL process
   ~$20

8. The number of incomplete ETL runs before your final run
   ~5 runs. But We only used very small amount of data and only 2 instances to do it

9. Discuss difficulties encountered
   There is some special characters in the Text of the tweets like '\n' and ',' which may have bad effect on the data loading after converting to csv. It was very annoying. We tried to use some censored word, which will never appear after the process in the text, to replace these two characters to cope with the problem.

   In very late stanes we found the correctness of the ETL is only ~ 76% after loading into MySQL. But it is already too late to correct it. And because our design of the csv file was not good enough, so we end up doing some string processing in the front end.

   Everytime we restart the MySQL server, the performance will be very very slow.

10. The size of the resulting database and reasoning
    ~ 58G ~200 million rows

11. The time required to backup the database on S3
    ~ 26 min

12. The size of S3 backup
    ~30GB

HBase:

13. The code for the ETL job
    Please refer to the attached CSV2HFiles.java

14. The programming model used for the ETL job and justification
    Use Map Reduce to get the HFile with Java API for trafering the csv file.
    Then configure the environment and load the data to Hbase.

15. The type of instances used and justification
    m3.large instances are used. For each instance, it should store informations and deal with info. So a larger instance is good for improving the QPS.

16. The number of instances used and justification
    A master node, 7 core node. It is used for EMR.

17. The spot cost for all instances used
    0.3

18. The execution time for the entire ETL process
    EMR: ~6.5 hour
    Data loading:

19. The overall cost of the ETL process
    ~$20

20. The number of incomplete ETL runs before your final run
    3

21. Discuss difficulties encountered
    Configuration of HBase, Hadoop map reduce, getting a HFiles.

22. The size of the resulting database and reasoning
    200 million rows in the database

23. The time required to backup the database on S3
    ~40 minutes

24. The size of S3 backup
    ~ 80GB

Questions

1. Describe a MySQL database and typical use cases.

MySQL is a relational database management system. Comparing to HBase, it may be used in relatively smaller amount of data (comparing to TB levels). And in situations that historical data is not so important.

2. Describe an HBase database and typical use cases.

Apache HBase™ is the Hadoop database, a distributed, scalable, big data store.Use Apache HBase when you need random, realtime read/write access to your Big Data.

3. What are the advantages and disadvantages of MySQL?

Advantages: data is very organized and structured. It can execute very complicated query within a reasonable time.

Disadvantage: Limited scalability. Performance highly depend on the index

4. What are the advantages and disadvantages of HBase?

Advantages: Handling data pretty fast, based on its property: distribution and extensibility. Can handle semi-structured data.

Disadvantages: hard to build the table.