**S15 15619 Project Phase 3 Report**

**Performance Data and Configurations**

|  |  |
| --- | --- |
| Best Configuration and | Results from the Live Test |
| Choice of backend (pick one) | HBase/MySQL |
| Number and type of instances | Live Test: |
| Cost per hour  (assume on-demand prices) | Live Test: |
| Queries Per Second (QPS) | INSERT HERE: (Q1,Q2,Q3,Q4,Q5, Q6)   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | |  | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | | score |  |  |  |  |  |  | | tput |  |  |  |  |  |  | | ltcy |  |  |  |  |  |  | | corr |  |  |  |  |  |  | | err |  |  |  |  |  |  | |
| Rank on the scoreboard: | Phase 1:  Phase 2:  Phase 3: |

**Team :**

**Members :**

**Rubric:**

**Each unanswered question = -10%**

**Each unsatisfactory answer = -5%**

**[Please provide an insightful, data-driven, colorful, chart/table-filled, and interesting final report. This is worth 20% of the grade for Phase 3. Use the report as a record of your progress, and then condense it before sharing it with us. Questions ending with “Why?” need evidence (not just logic)]**

**Task 1: Front end (you may/should copy answers from your earlier report-- each report should form a comprehensive account of your experiences building a cloud system. Please try to add more depth and cite references for your earlier answers)**

**Questions**

1. Which front end framework did you use? Explain why you used this solution. [Provide a small table of special properties that this framework/platform provides].
2. Did you change your framework after Phase 1 or Phase 2? Why or why not?
3. Explain your choice of instance type and numbers for your front end system.
4. Explain any special configurations of your front end system.
5. Did you use an ELB for the front-end? Why, or why not? Condense your experience with ELB in the next few sentences. Talk about load-balancing in general and why it matters in the cloud.
6. Did you explore any alternatives to ELB? List a few of these alternatives. What did you finally decide to use? (if possible) Provide some graphs comparing performance between different types of systems.
7. Did you automate your front-end instance? If yes, how? If no, why not?
8. Did you use any form of monitoring on your front-end? Why or why not? If you did, show us a capture of your monitoring during the Live Test. Else, try to provide CloudWatch logs of your Live Test in terms of memory, CPU, disk and network utilization. Demarcate each query clearly in the submitted image capture.
9. What was the cost to develop the front end system?
10. What are the best reference URLs (or books) that you found for your front-end? Provide at least 3.

[Please submit the code for the frontend in your ZIP file]

**Task 2: Back end (database)**

**Questions**

1. Which DB system did you choose in Phase 3? Why? Would any different queries for Q5 and Q6 have influenced you to choose the other DB?
2. Describe your schema. Explain your schema design decisions. Would your design be different if you were not using this database? How many iterations did your schema design require? Also mention any other design ideas you had, and why you chose this one? Answers backed by evidence (actual test results and bar charts) **are required**.
3. What was the most expensive operation / biggest problem with each DB that you had to resolve for each query? Why does this problem exist in this DB? How did you resolve it? Plot a chart showing the improvements with time.
4. Explain (briefly) **the theory** behind (at least) 11 performance optimization techniques for databases. How are each of these implemented in MySQL? How are each of these implemented in HBase? Which optimizations only exist in one type of DB? How can you simulate that optimization in the other (or if you cannot, why not)? Use your own words (paraphrase, this document goes through plagiarism detection software). **(this question is worth 20 points if not answered in detail)**
5. Plot a graph showing results with/without each individual optimization that you used. Extremely impressive will be a timeline of rps v/s submission id (mentioning which optimization was in use at that time).
6. Would your design work if your web service also implemented insert/update (PUT) requests? Why or why not?
7. Which API/driver did you use to connect to the backend? Why? What were the other alternatives that you tried? What changed from Phase 1? From Phase 2?
8. Can you quantify the speed differential (in terms of rps or Mbps) between reading from disk versus reading from memory? Did you attempt to maximize your usage of RAM to store your tables? How much (in % terms) of your memory could you use to respond to queries?
9. Did you use separate tables for Q2-Q6? How can you consolidate your tables to reduce memory usage?
10. What are the flaws you have seen in both DBs?
11. How did you profile the backend? If not, why not? Given a typical request-response for each query (Q2-Q6) what percentage of the overall latency is due to:
    1. Load Generator to Load Balancer (if any, else merge with b.)
    2. Load Balancer to Web Service
    3. Parsing request
    4. Web Service to DB
    5. At DB (execution)
    6. DB to Web Service
    7. Parsing DB response
    8. Web Service to LB
    9. LB to LG

**How did you measure this**? A 9x5 (Q2 to Q6) table is one possible representation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Q2** | **Q3** | **Q4** | **Q5** | **Q6** |
| **Load Generator to Load Balancer (if any)** |  |  |  |  |  |
| **Load Balancer to Web Service** |  |  |  |  |  |
| **Parsing request** |  |  |  |  |  |
| **Web Service to DB** |  |  |  |  |  |
| **At DB (execution)** |  |  |  |  |  |
| **DB to Web Service** |  |  |  |  |  |
| **Parsing DB response** |  |  |  |  |  |
| **Web Service to LB** |  |  |  |  |  |
| **LB to LG** |  |  |  |  |  |
|  | 100% | 100% | 100% | 100% | 100% |

1. What was the cost to develop your back end system?
2. What were the best resources (online or otherwise) that you found. Answer for HBase, MySQL and any other relevant resources.

[Please submit the code for the backend in your ZIP file]

**Task 3: ETL**

1. For each query, write about:
   1. The programming model used for the ETL job and justification
   2. The number and type of instances used and justification
   3. The spot cost for all instances used
   4. The execution time for the entire ETL process
   5. The overall cost of the ETL process
   6. The number of incomplete ETL runs before your final run
   7. Discuss difficulties encountered
   8. The size of the resulting database and reasoning
   9. The size of the backup
2. What are the most effective ways to speed up ETL? How did you optimize writing to your backend? Did you make any changes to your tables after writing the data? How long does each load take?
3. Did you use EMR? Streaming or non-streaming? Which approach would be faster and why?
4. Did you use an external tool to load the data? Which one? Why?
5. Which database has been easier to load (MySQL or HBase)? Why? Has your answer changed in the last four weeks?

[Please submit the code for the ETL job in your ZIP file]

**General Questions**

1. Would your design work as well if the quantity of data would double? What if it was 10 times larger? Why or why not?
2. Did you attempt to generate load on your own? If yes, how? And why?
3. Describe an alternative design to your system that you wish you had time to try.
4. Which was/were the toughest roadblock(s) faced in Phase 3?
5. Did you do something unique (any cool optimization/trick/hack) that you would like to share with the class?

**Bonus Questions**

1. Draw the Data Dependency DAG of fan-out requests (see P3.3 recitation [slide 10](http://www.cs.cmu.edu/~msakr/15619-s15/recitations/Recitation09.pdf) for an example of this graph) for the following sequence of events:
   1. Open the bonus page
   2. Search #beautiful between 2014-01-01 and 2014-12-31
   3. Click on the first 10 tweets
   4. Click on the first 10 users
2. Did you use any parallelization to speed up this data fetching at the front-end? Did you use any front-end web technique to make it appear faster for a visitor to the web page?
3. Show the network graph from your browser when 1a-1d are performed.