

15-319 / 15-619

Cloud Computing

Recitation 11

Tuesday, Nov 5, 2019

Overview

- **Last week's reflection**

- Project 4.1
- Quiz 9
- Team Project Phase 2 released

- **This week's schedule**

- Unit 5 - Modules 19 & 20
 - Quiz 10
- Team Project, Phase 2, Queries, 1, 2, 3
- Team Project, live test
 - HBase
 - MySQL

P4.1 Reflection

- Programming in Scala and Spark
- Understanding the differences between processing data with MapReduce and Spark
- Exploring Twitter social data with RDD and SparkSQL APIs
- Implementing an iterative processing algorithm - pagerank - on a large dataset
- Utilizing the Spark Web UI to monitor a Spark job and identify performance bottlenecks
- Tuning a Spark program to optimize for time
- Running the PageRank application on Azure Databricks to compare performance

P4.1 Reflection

- Common Issues
 - Handling dangling nodes in the graph
 - Tuning the cluster for better performance.
 - Long running jobs
 - Reduce the amount of data shuffling
- Takeaways
 - Some approaches to implementing pagerank are more efficient than others
 - The Spark Web UI is a useful visualization tool
 - Databricks offers optimized version of Spark providing better performance than HDInsight.

Modules to Read

- UNIT 5: Distributed Programming and Analytics Engines for the Cloud
 - Module 18: Introduction to Distributed Programming for the Cloud
 - Module 19: Distributed Analytics Engines for the Cloud: MapReduce
 - Module 20: Distributed Analytics Engines for the Cloud: Spark



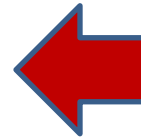
TEAM PROJECT

Twitter Data Analytics



Team Project

- Phase 1:
 - Q1
 - Q2 (MySQL AND HBase)
- Phase 2
 - Q1
 - Q2 & Q3 (MySQL AND HBase)
- Phase 3
 - Q1
 - Q2 & Q3 (MySQL OR HBase)



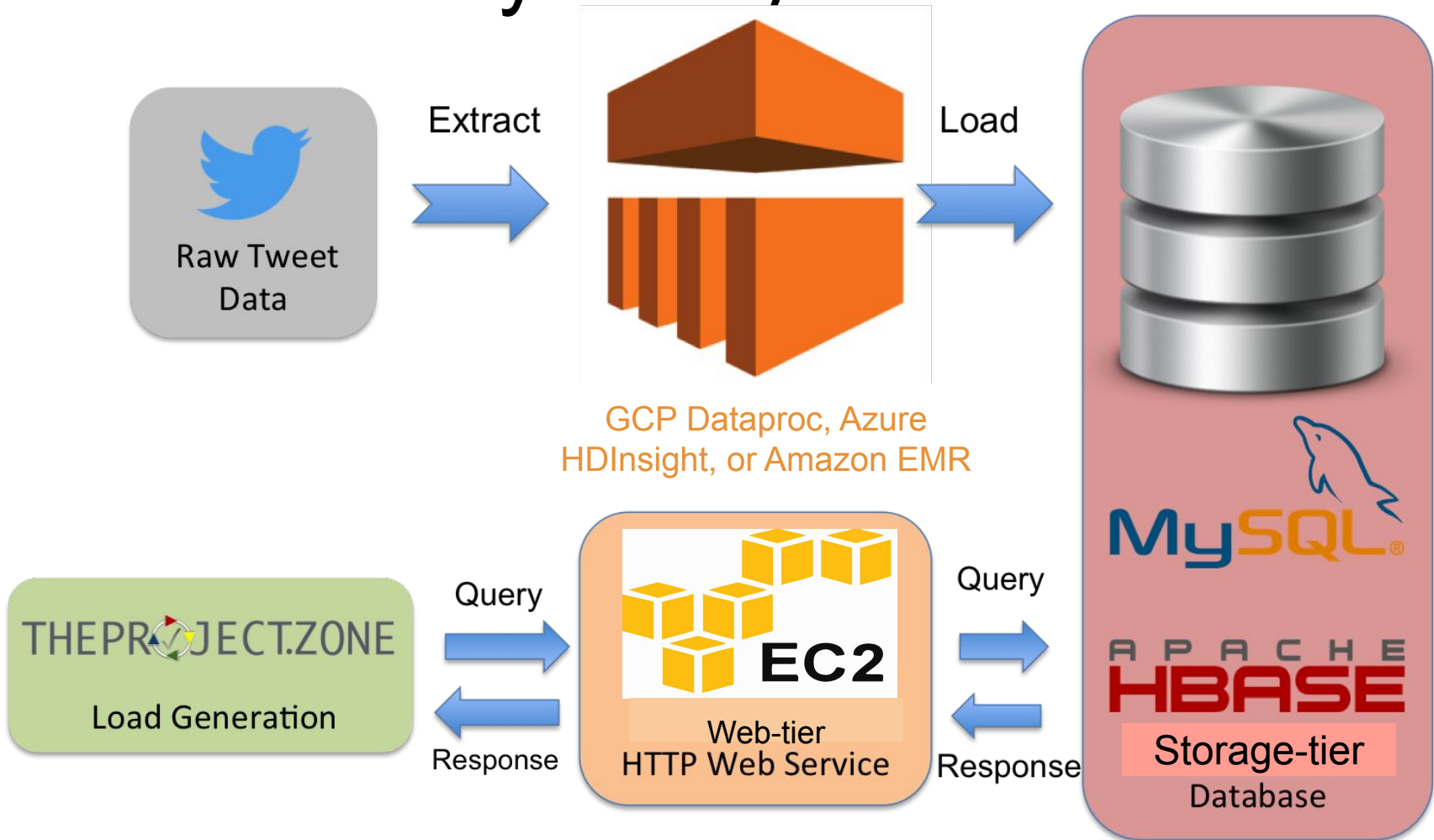
Team Project

Twitter Analytics Web Service

- Given ~1TB of Twitter data
- Build a performant web service to analyze tweets
- Explore web frameworks
- Explore and optimize database systems



Twitter Analytics System Architecture



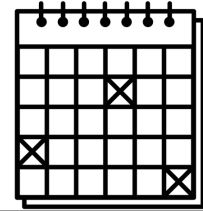
- Web server architectures
- Dealing with large scale real world tweet data
- HBase and MySQL optimization



Team Project Deadlines

- Phase 2 milestones:
 - Phase 2, Live test: on **Sunday, November 10**
 - HBase:
 - Q1/Q2/Q3/mixed
 - MySQL:
 - Q1/Q2/Q3/mixed
 - Phase 2, code, scripts and report:
 - due on **Tuesday, November 12**

Team Project Time Table



Phase (and query due)	Start	Deadlines	Code and Report Due
Phase 1 <ul style="list-style-type: none"> Q1, Q2 	Monday 10/07/2019 00:00:00 ET	Checkpoint 1, Report: Sunday 10/13/2019 23:59:59 ET Checkpoint 2, Q1: Sunday 10/20/2019 23:59:59 ET Phase 1, Q2: Sunday 10/27/2019 23:59:59 ET	Phase 1: Tuesday 10/29/2019 23:59:59 ET
Phase 2 <ul style="list-style-type: none"> Q1, Q2, Q3 	Monday 10/28/2019 00:00:00 ET	Sunday 11/10/2019 15:59:59 ET	
Phase 2 Live Test (Hbase AND MySQL) <ul style="list-style-type: none"> Q1, Q2, Q3 	Sunday 11/10/2019 17:00:00 ET	Sunday 11/10/2019 23:59:59 ET	Tuesday 11/12/2019 23:59:59 ET
Phase 3 <ul style="list-style-type: none"> Q1, Q2, Q3 (Managed services) 	Monday 11/11/2019 00:00:00 ET	Sunday 11/24/2019 15:59:59 ET	
Phase 3 Live Test <ul style="list-style-type: none"> Q1, Q2, Q3 (Managed services) 	Sunday 11/24/2019 17:00:00 ET	Sunday 11/24/2019 23:59:59 ET	Tuesday 11/26/2019 23:59:59 ET

Live Test Schedule - setup

Submit DNS for Live Test

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. Last chance to update your DNS for the HBase Live Test
5:33 pm - 5:34 pm	MySQL DNS Validation	Validate your MySQL DNS. Last chance to update your DNS for the MySQL Live Test

Live Test Schedule - HBase

HBase Live Test

Information

Time	Value	Target	Weight
6:00 pm - 6:25 pm	Warm-up (Q1 only)	0	0%
6:25 pm - 6:50 pm	Q1	35000	6%
6:50 pm - 7:15 pm	Q2	10000	10%
7:15 pm - 7:40 pm	Q3	2000	10%
7:40 pm - 8:05 pm	Mixed Reads(Q1,Q2,Q3)	9000/2500/500	4+5+5 = 14%

Half-time Break

Information

Time	Value
8:05 pm - 8:30 pm	Time to relax and prepare for the MySQL Live Test

Live Test Schedule - MySQL

MySQL Live Test

Information

Time	Value	Target	Weight
8:30 pm - 8:55 pm	Warm-up (Q1 only)	0	0%
8:55 pm - 9:20 pm	Q1	35000	6%
9:20 pm - 9:45 pm	Q2	10000	10%
9:45 pm - 10:10 pm	Q3	2000	10%
10:10 pm - 10:35 pm	Mixed Reads(Q1,Q2,Q3)	9000/2500/500	4+5+5 = 14%

Budget Reminder

- Your team has a total AWS budget of **\$50** for Phase 2
- Your web service should cost \leq **\$0.89/hour**, including:
 - EC2
 - We evaluate your cost using the [On-Demand Pricing](#) towards **\$0.89/hour** even if you use spot instances.
 - EBS & ELB
 - Ignore data transfer and EMR cost
- Phase 2 - Live Test Targets:
 - Query 1 - 35000 rps
 - Query 2 - 10000 rps (for both MySQL and HBase)
 - Query 3 - 3000 rps (for both MySQL and HBase)
 - Mixed - 9000/2500/500 rps (for both MySQL and HBase)

Phase 2, Query 3

- **Problem Statement**

- Given a time range and a user id range, which tweets have the most **impact** and what are the **topic words**?

- Impact score and topic words (see the write up for details)

- Impact of tweets: Which tweet is “important”? Calculate using the effective word count, favorite count retweet count and follower count.
- Topic words: In this given range, what words could be viewed as a “topic”? Done using TF-IDF.

- Request/Response Format

- Request: Time range, uid range, #words, #tweets
- Response: List of topic words with their topic score, as well as a list of tweets (after censoring)

Phase 2, Query 3 FAQs

Question 1: How to calculate the topic score?

For word w in the given range of tweets, calculate:

- Calculate the Term Frequency of word w in tweet $t^{(i)}$
- Calculate Inverse Document Frequency for word w
- Calculate Impact Score of each tweet
- Topic Score for word $w =$
$$\sum_i^n TF(w, t^{(i)}) \cdot IDF(w) \cdot \ln(\text{Impact}(t^{(i)}) + 1),$$

for n tweets in time and uid range

Phase 2, Query 3 FAQs

Question 2: When to censor? When to exclude stop words?

- Censor in the Web Tier or during ETL. It is your own choice.
 - If you censor in ETL, consider the problem it brings to calculating the topic word scores (two different words might look the same after censoring).
- You should count stop words when counting the total words for each tweet in order to calculate the topic score.
- Exclude stop words when calculating the impact score and selecting topic words.

Hints

- Completely understand every AssessMe question
- Completely understand the definition of a word. This is different for text censoring and calculating scores.
- A query contains two ranges. Log some requests to get an idea on the range of user_id and timestamps.
- Optimization is time-consuming. Before ETL, please
 - Think about your schema design
 - Think about your database configuration

Hints

- For HBase, you're not restricted to just 1 master node. The two sample setups below are both permitted.
 - 1 x (1 master + 5 slaves)
 - 2 x (1 master + 2 slaves)
- Understand and keep an eye on
 - **EC2 CPU Credits** and burstable performance
 - **EBS volume I/O Credits** and Burst Performance

Hints for the live test

- The request pattern will differ for Phase 2 submission test and the live test so your solution should handle all types of load.
- Monitor your system during the live test to recover in case of a system crash.
- Be prepared with your monitoring consoles setup
- Lookup what commands you can use to learn about the aspects of your web service health.
- Your Phase 2 budget should take into account the cost for the live test.
- Take cloudwatch snapshots.

Warning

- NEVER open all ports to the public (0.0.0.0) when using instances on a public cloud.
- For your purposes, you likely only need to open port 80 to the public. Port 22 should be open only to your public IPs.
- Port 3306 (for MySQL) and HBase ports should be open only to cluster members if necessary.

Upcoming Deadlines

- P4.1 Spark
 - **Code review - this week**
- Quiz 10
 - **Due: 11/08/2019 11:59 PM Pittsburgh**
- Team Project : Phase 2
 - **Live-test due: 11/10/2019 3:59 PM Pittsburgh**
 - **Code and report due: 11/12/2019 11:59 PM Pittsburgh**

Questions?

