15-319 / 15-619
Cloud Computing

Recitation 14

Apr 25, 2017

Overview

- **Last week**
    - P4.2, Iterative Processing with Spark
    - Team Project, Phase 3
- **This week's schedule**
    - Quiz 12
    - **Twitter Analytics**
        - Team Project, Phase 3, Live Test

Reminders

- Tag all resources and monitor AWS expenses regularly
  - Check your bill (Cost Explorer > filter by tags)
  - Use the team AWS account and tag the Team Project resources carefully
- Piazza Guidelines
  - Please tag your questions appropriately
  - Search for an existing answer first
- Provide clean, modular and well documented code
  - Large penalties for not doing so
  - **Double check** that your code is submitted! (Verify by clicking "download" from TPZ from the submissions page)
- Utilize Office Hours
  - We are here to help (but not to give solutions)

Module 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
  - Module 21: Distributed Analytics Engines for the Cloud: GraphLab
  - Module 22: Message Queues and Stream Processing

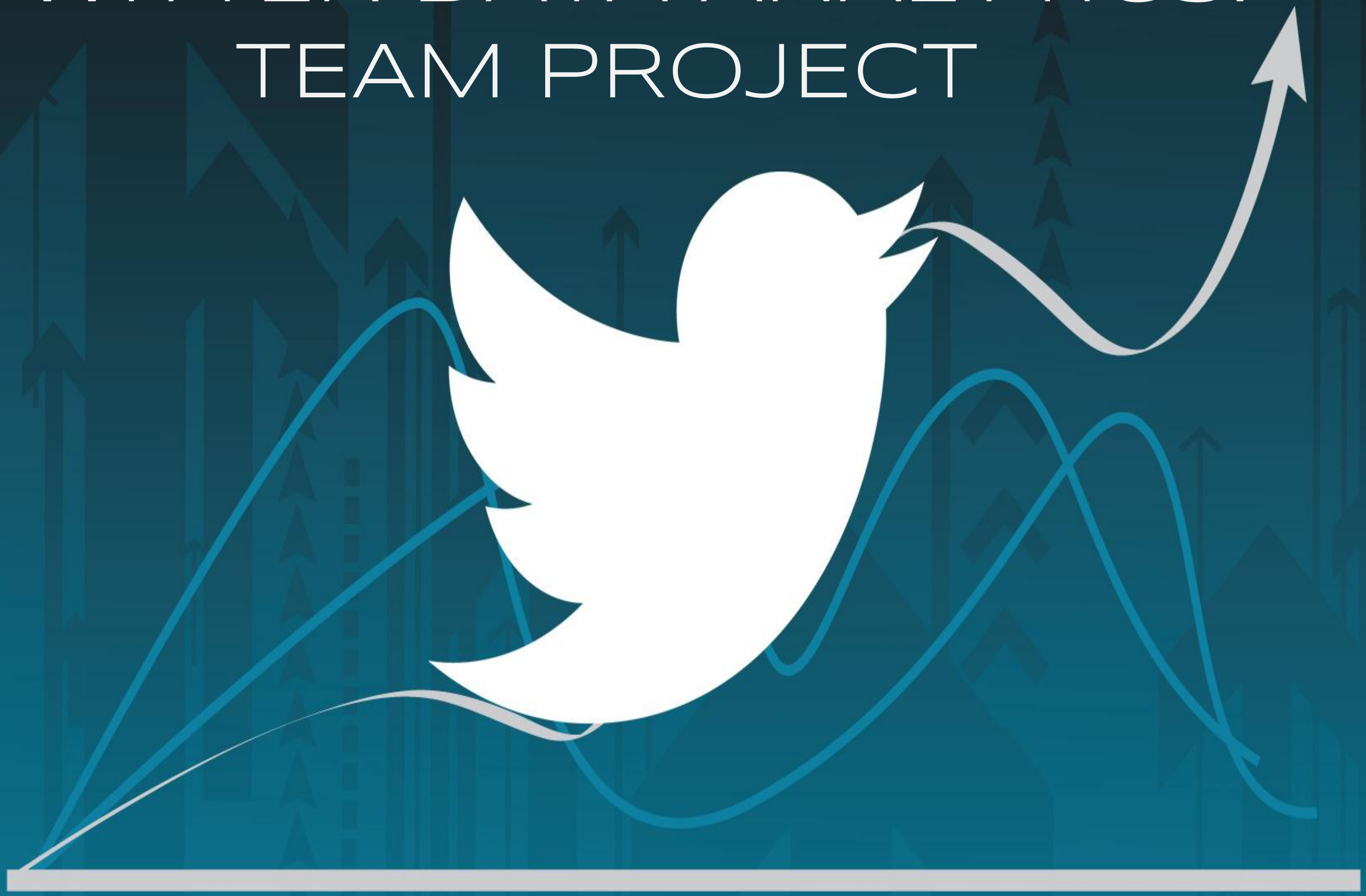
P4.2 Reflections

Please complete the project survey here

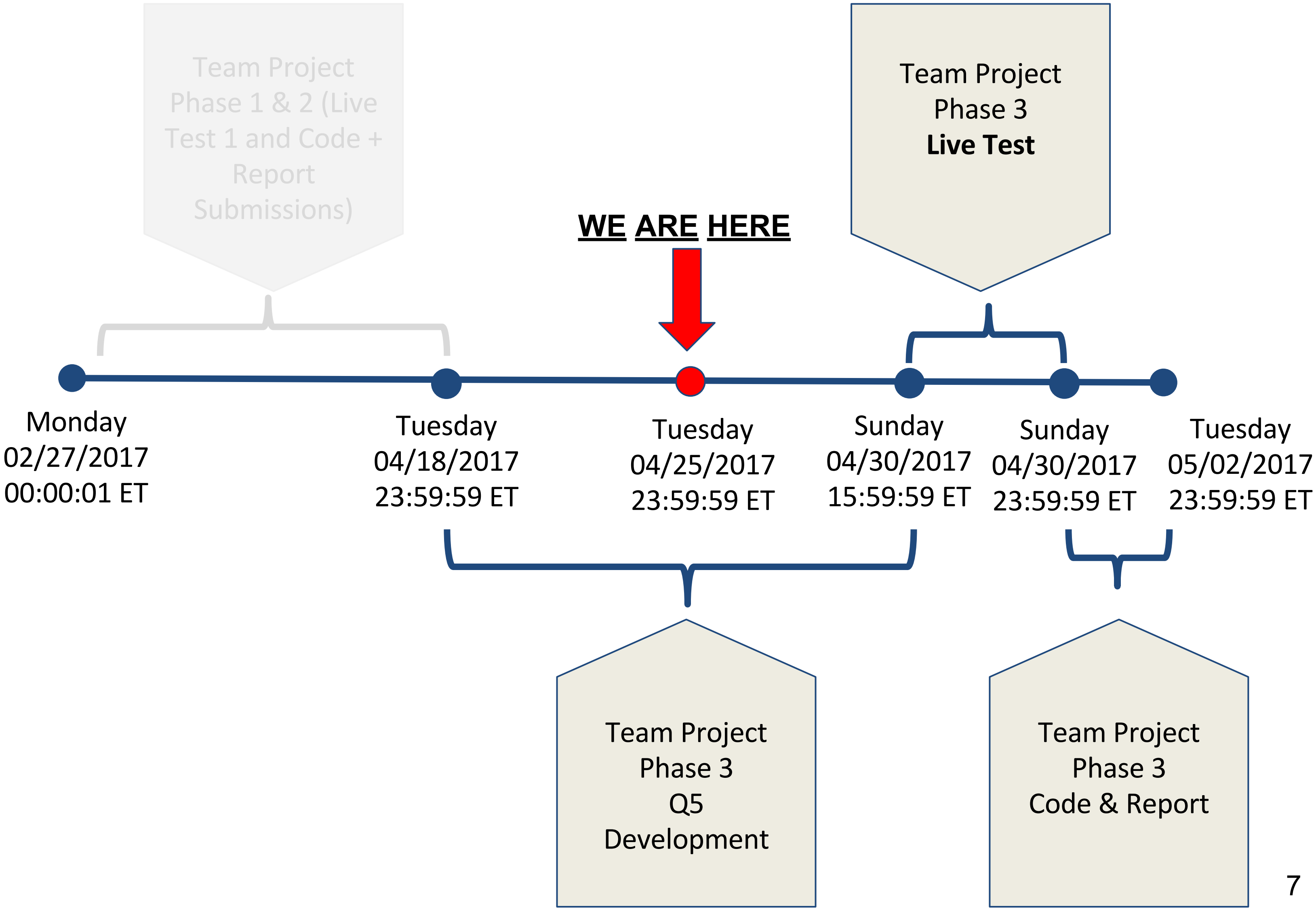
What you learned:

- Basic graph exploration (vertices, edges, degree) on Spark.
- PageRank algorithm and iterative batch processing on Spark.
- Graph data exploration on GraphX.
- Monitor and tune your Spark program.
- Programming in Scala on Spark.

TWITTER DATA ANALYTICS: TEAM PROJECT



Team Project, Phase 3 Deadlines



Team Project Phase 1 & 2 (Live Test 1 and Code + Report Submissions)

WE ARE HERE

Team Project Phase 3 **Live Test**

Monday 02/27/2017 00:00:01 ET

Tuesday 04/18/2017 23:59:59 ET

Tuesday 04/25/2017 23:59:59 ET

Sunday 04/30/2017 15:59:59 ET

Sunday 04/30/2017 23:59:59 ET

Tuesday 05/02/2017 23:59:59 ET

Team Project Phase 3 Q5 Development

Team Project Phase 3 Code & Report

Team Project Time Table

Phase (and query due)	Start	Deadline	Code and Report Due
Phase 1 ● Q1, Q2	Monday 10/10/2016 00:00:01 EST	Sunday 10/30/2016 23:59:59 ET	**Tuesday 11/01/2016 23:59:59 ET**
Phase 2 ● Q1, Q2, Q3, Q4	Monday 10/31/2016 00:00:01 ET	Sunday 11/13/2016 **15:59:59 ET**	
Phase 2 Live Test (Hbase/MySQL) ● Q1, Q2, Q3, Q4	Sunday 11/13/2016 18:00:01 ET	Sunday 11/13/2016 23:59:59 ET	Tuesday 04/18/2017 23:59:59 ET
Phase 3 ● Q1, Q2, Q3, Q4, Q5	Monday 04/17/2017 00:00:01 ET	**Sunday 04/30/2017 15:59:59 ET**	
Phase 3 Live Test ● Q1, Q2, Q3, Q4, Q5	**Sunday 04/30/2017 18:00:01 ET**	**Sunday 04/30/2017 23:59:59 ET**	Tuesday 05/02/2017 23:59:59 ET

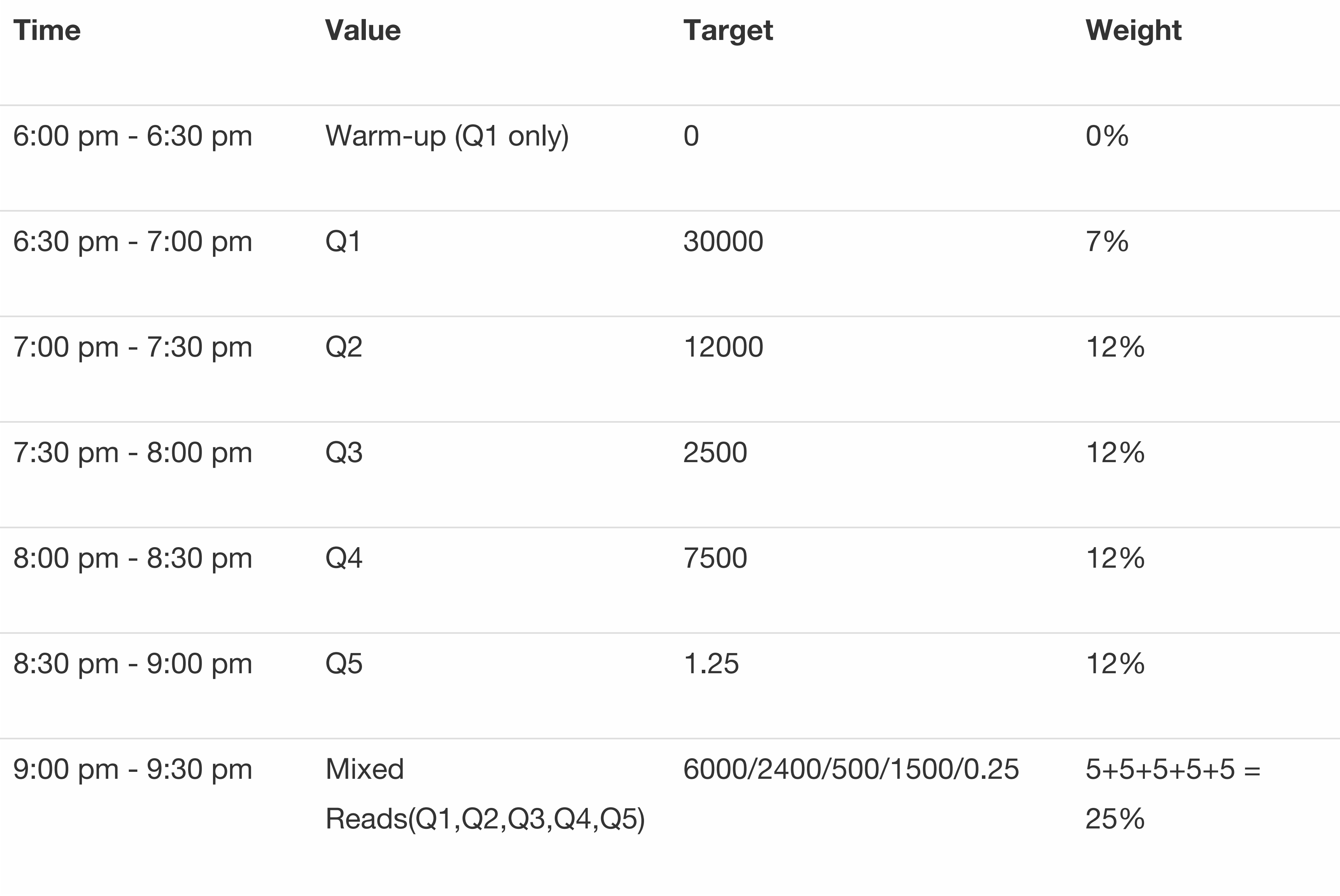
Team Project, Phase 3

- Budget on AWS for Phase 3 is $50 (including the Live Test), if you spend more than $75 you will receive a 100% penalty.
- You have a $0.88/hr budget in the Live Test. You will receive (X-0.88)*2% penalty if you spend X dollars > $0.88. The hourly cost includes:
    - **EC2**
        - We evaluate your cost using the On-Demand Pricing towards **$0.88/hour** even if you use spot instances.
    - **EBS & ELB**
    - Ignore data transfer and EMR cost
- We encourage you to use Azure and GCP for their ETL jobs in order to save your AWS budget for development, testing and the Live Test.

Team Project, Phase 3

- Remember to tag your HBase(Key: **teambackend** Value: **hbase**) or MySQL(Key: **teambackend** and Value: **mysql**) cluster.
- You must submit your DNS for the Live Test before 4 pm, Sunday Apr 30th.
- You must use the same cluster for all queries.
- Do not launch other testing instances during the Live Test, or else we will count them towards your hourly budget.
- We encourage you to use on-demand instances for the Live Test or else you run the risk of your instances being shut down unexpectedly.
- Leave enough budget for the Live Test.

Phase 3, Live Test Schedule



Time	Value	Target	Weight
6:00 pm - 6:30 pm	Warm-up (Q1 only)	0	0%
6:30 pm - 7:00 pm	Q1	30000	7%
7:00 pm - 7:30 pm	Q2	12000	12%
7:30 pm - 8:00 pm	Q3	2500	12%
8:00 pm - 8:30 pm	Q4	7500	12%
8:30 pm - 9:00 pm	Q5	1.25	12%
9:00 pm - 9:30 pm	Mixed Reads(Q1,Q2,Q3,Q4,Q5)	6000/2400/500/1500/0.25	5+5+5+5+5 = 25%

Query 5, Graph Query

- Finding the shortest path from User A to B in the mention graph, built from the original dataset.
- Useful fields in a JSON tweet when doing extraction:
  - [“id_str”]
  - [“user”][“id_str”]
  - [“entities”][“user_mentions”]
  - You should ignore the tweet if any of the above fields is invalid.
- Target RPS is **1.25**
- Warning: You must get a correctness of over 80% to get a score.

Team Project, General Hints

- Identify the bottlenecks using fine-grained profiling.
- Do not cache naively.
- Review what we have learned in previous project modules
    - Scale out
    - Load balancing (Are requests balanced)
    - Replication and sharding
- Look at the feedback of your previous reports!

Team Project, Query 4 Hints

- Start with one machine if you are not sure your concurrency model is correct pay attention to scalability.
- Forwarding mechanism or non-forwarding mechanism
  - You may want to use a L7 load balancer.
- May need many connections at the same time, in the case of out of order sequence numbers.
- Consider batch writes. Write and read are exclusive due to the consistency model.

Team Project, Phase 3 Deadlines

- Phase 3 Development
  - **Submission deadline 15:59 ET (Pittsburgh) Sunday 04/30**
    - **Live Test from 6 PM to 10 PM ET**
  - Fix Q1 - Q4 if you did not go well
  - Query 5
  - Phase 3 counts for **50%** of the Team Project grade!
- Phase 3 Report
  - **Submission 23:59:59 ET (Pittsburgh) Tuesday 05/02**
  - Explain in detail the strategies you used
  - Difficulties you encountered even if you didn't get a good score

Upcoming Deadlines

- Quiz 12
  - Due: **04/28**/2017 11:59 PM Pittsburgh
- Team Project : Phase 3
  - Live-test due: **04/30**/2017 **3:59 PM** Pittsburgh
  - Code and report due: **05/02**/2017 11:59 PM Pittsburgh

Questions?