CS15-319 / 15-619
Cloud Computing

Recitation 15

15619 Project Review

Apr 29th & May 1st 2014

http://www.qatar.cmu.edu/~msakr/15619-s14/

Congratulations!

- 15619 Project is done!
- We are proud of you!
  - 30 teams got 70+ in Phase 3 live test
  - 3 full score teams
  - Many fantastic designs
- Let's wrap everything up and discuss some take-home lessons

15619 Project Summary

- 15619 project is different
  - An open project
    - Only query, budgets and business requirements given
    - Very little hand-holding
    - That's the real world!

  - No well established solution
  - A project on the cloud
    - How to deal with uncertainty…
    - Very expensive…

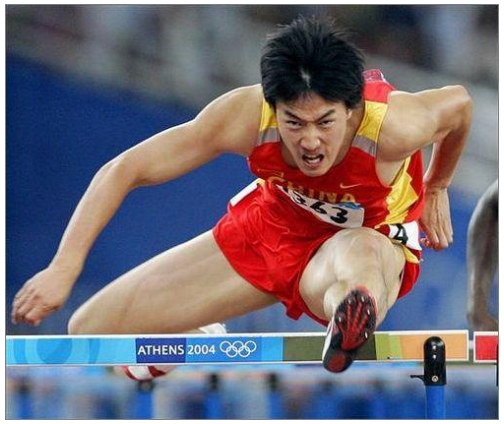
15619 Project Summary

Project 1
Sequential Analysis
MapReduce with Hadoop Streaming

Project 2
ELB
Autoscaling

Project 3
MySQL & HBase
Horizontal / Vertical scaling

15619 Project
Q1: front end
Q2: single tweet
Q3: Retweeters
Q4: Tweet Texts
Q5: Place + Time range query
Q6: User ID range



Project 4
MapReduce
N-Gram & Language Statistical Model

15619 Project Summary

- What else besides the bricks (project modules)
  - System architecture design
  - Database schema design
  - Bottleneck analysis
  - Performance optimization
  - Deal with uncertainty
    - E.g. Which availability zone should we use?

Ability to handle a real world project on the cloud!

Query Designs Purpose

- Q1:Front End
- Q2: Big number of records
- Q3: Small number of records
- Q4: Large query
- Q5: ETL focused
- Q6: Database table design + ETL

Front end

- Connection Pool
- Some settings
  - Memory size
  - Thread numbers
- Usually, less post-query processing is preferred
  - Pre-processing the data
    - E.g. store sorted answer in DB
- …

MySQL

- MySQL index (Could be tricky)
  - Joint index?
  - Integer index
- Number of records to check matters
  - “Explain” the SQL statement
    - Scan is not free

HBase

- Row key design
  - Byte stream comparison
  - Shorter key perform better
- Get is better than Scan
- Load balance (region/split settings)

15619 Project Notes

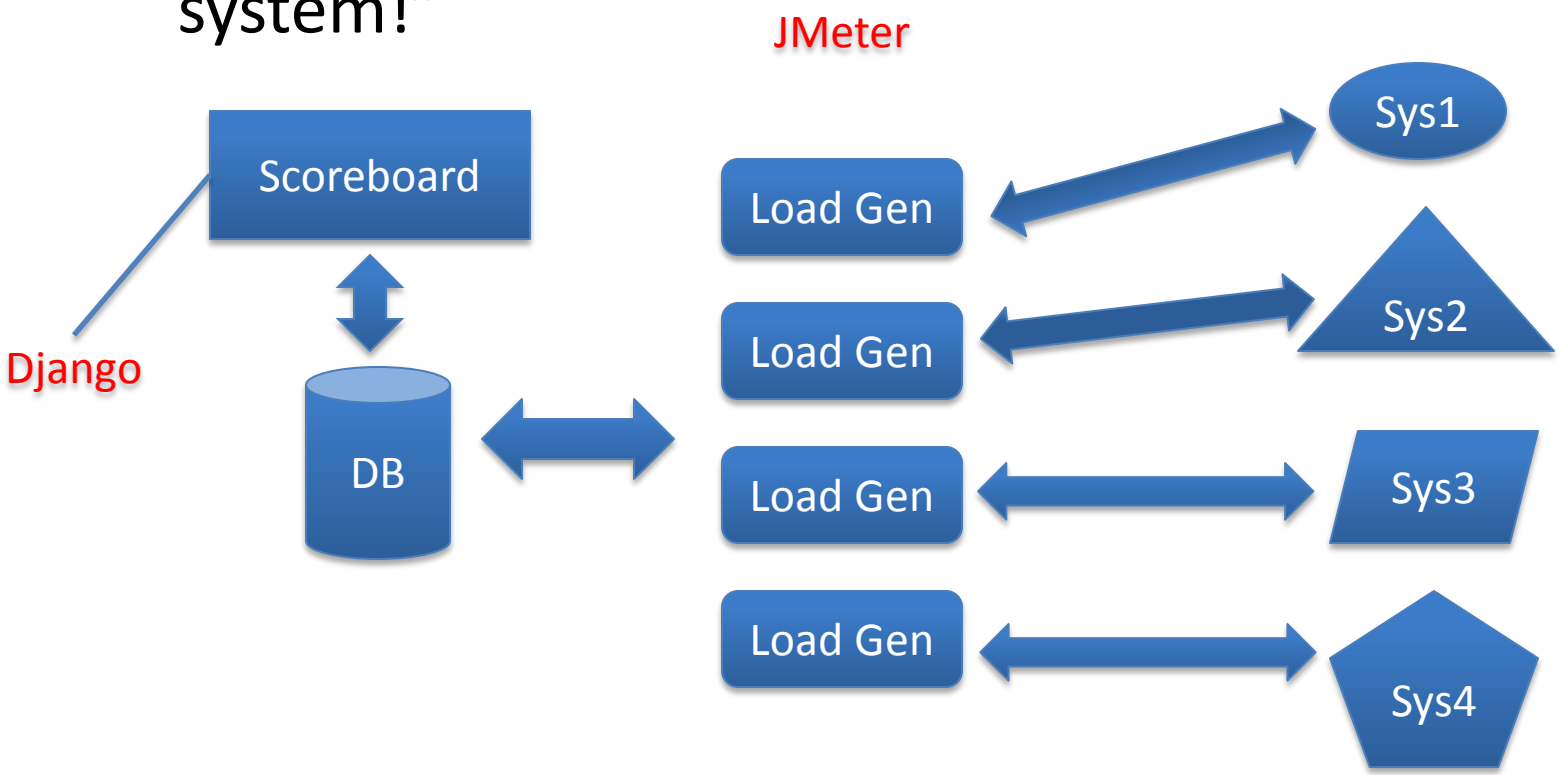
- Some misunderstandings
  - "I got nothing with more instances after ELB. It must haven't been warmed up yet!"
    - **Max_Throughput = 1 / latency * #Threads**
    - We have 50 threads for most queries (not a big number)
    - **Resp. Time = Transmission Time + Propagation Time + Queuing Time + Serving Time**
    - Adding ELB will benefit the performance by reducing the queuing time because it has more instances serving the requests. However, it does not improve other elements of time. Maybe even making them become longer.

15619 Project Notes

- Some misunderstandings
  - "We did way better in regular test than in live test, there must be something wrong on AWS!"
    - It is possible that something on AWS caused the variation
    - But not necessarily so
      - Exact same system could have totally different performance because of the duration of the test.
      - The sequence of the test queries matters
      - We also encountered other cases…

15619 Project Summary

- Some misunderstandings
  - "Your little Django is the bottleneck of testing system!"



15619 Project

- Through practice, you have learnt
  - System architecture design
  - Database schema design
  - Bottleneck analysis
  - Performance optimization
  - Deal with uncertainty

Ability to work on a real world cloud project!

Questions?

Upcoming Deadlines

- Unit 5:

UNIT 5: Distributed Programming and Analytics Engines for the Cloud		
Module 16: Introduction to Distributed Programming for the Cloud		
Module 17: Distributed Analytics Engines for the Cloud: MapReduce		
Module 18: Distributed Analytics Engines for the Cloud: Pregel		
Module 19: Distributed Analytics Engines for the Cloud: GraphLab		
Quiz 5: Distributed Programming and Analytics Engines for the Cloud	Checkpoint	Available Now Due 05/1/14 11:59 PM