Final Review

DS 5110: Big Data Systems

Spring 2025

Yue Cheng

Final exam

- Thursday, May 8, 9:00 am – 11:00 am
  - Open book, open notes

- Covering six topics from Lec 8 to Lec 19
  - Spark RDD
  - Ray
  - Cloud computing
  - Serverless computing
  - S3 and HDD
  - Dynamo and consistent hashing

Logistics

- The exam will be remote + synchronous over `gradescope`
- The exam sheet will be available on `gradescope` at 9 am
- You should work directly on `gradescope`
- Submission closes at 11:15 am (a grace period of 15 minutes for submission)

Theme 1: Big data systems

Spark

- Motivation
- Transformations and actions
- The use of `.persist()` in iterative applications like PageRank

Ray

- Ray's programming APIs
  - Tasks: executing stateless code
  - Actors: stateful

- What apps Ray can support
  - Generic parallel data processing → You can implement a Spark atop Ray
  - Complex ML/AI workflows: RL, pretraining/inference, etc.

Theme 2: Cloud computing

Cloud computing

- Infrastructure-as-a-Service (IaaS)

- Cloud pricing: “Pay-as-you-go”
  - What’s the problem?
  - Challenges of performing strategic resource planning

- Incentivizing tenants to use less during peak hours and use more in off-peak periods
  - On-demand VMs
  - Spot VMs

Serverless computing

- Function-as-a-Service (FaaS)

- How AWS Lambda works
  - Lambda invocation/triggering
  - Provider provisions Lambda function instance(s)
    - Fast path: Hot/warm start 🔥
    - Slow path: Cold start ❄️
  - Lambda function starts execution → billing begins
  - Lambda function terminates and → billing stops

- Desirable properties of today's FaaS
  - Autoscaling and scaling down to zero
  - Closer to "pay-per-use"
- Limitations of today's FaaS

Theme 3: Cloud storage systems

https://www.allthingsdistributed.com/2023/07/building-and-operating-a-pretty-big-storage-system.html

AWS S3

- S3 relies on HDDs (hard disk drives) for cost-effective storage

- HDD's working mechanism
  - Performance model: $L_{I/O} = L_{seek} + L_{rotate} + L_{transfer}$
  - Entire **seek** often takes 4 - 10ms
  - **Rotation** per minute (RPM): 7,200 RPM is common
  - **Transfer** is relatively faster compared to other two phases

- S3 workloads can be spiky, so data placement is crucial for performance

Amazon Dynamo

- Dynamo uses consistent hashing for data partitioning

- How consistent hashing works
    - Ring-shaped name space
    - Token maps of nodes
    - Virtual nodes
    - How to support replication

Putting it all together

- Theme 1: Big data systems
- Theme 2: Cloud computing
- Theme 3: Cloud storage systems

Question types

- Multi-choice questions

- True or false questions

- Problem solving

Thank you all for a great semester!

- Still, one last guest lecture next Tuesday
  - Hugging Face Xet

- Wish you all the best!

TIME WELL SPENT™
by Tom Fishburne
SEE, I TOLD YOU THAT BIG DATA WAS TOO SCARY
BIG DATA
KRONOS
Workforce Innovation That Works™
© 2012
KRONOS.COM/TIME WELLSPENT

Quiz 6

- Please fill out the informal teaching evaluation form
  - Anonymous, not mandatory, but with extra credit

- Please fill out the SET (Student Experience of Teaching) form