15-319 / 15-619
Cloud Computing

Overview 11

November 9th, 2021

Reflection of Last Week

- **OLI, Unit 5**
  - 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
- **Quiz 9 (OLI Module 18)**
- **Project 5 - Stream Processing with Kafka and Samza**
- **Team Project Phase 2**

This Week

- **OLI, Unit 5**
  - 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

- **Quiz 10 (OLI Module 19 and 20)**
  - Due on **Friday**, November 12th, 2021, 11:59PM ET

- **Project 5 - Stream Processing with Kafka and Samza**
  - Due on **Sunday**, November 14th, 2021, 11:59PM ET

- **Team Project Phase 2 Report**
  - Due on **Tuesday**, November 9th, 2021, 11:59PM ET
- **Team Project Phase 3**
  - Started this week!

TEAM PROJECT
Twitter Data Analytics

Top Microservice 1 Teams

Submitter	Effective RPS
ByteWave	153997.93
MidAutumn	142963.63
Absolute	120107.93

Top Microservice 2 Teams

Submitter	Effective RPS
CloudyWithAChanceOf Meatballs	46630.16
NoeatNosleep2021	42195.53
CloudComputers	41569.32

Top Microservice 3 Teams

Submitter	Effective RPS
CloudK	18891.49
Absolute	12680.93
InvolutionKings	9694.57
HeadInTheClouds	9595.92

Team Project Time Table
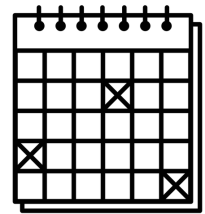
Task	Timeline
Phase 2 Report	● Due: Tuesday, 11/9/2021 11:59 PM EST
Phase 3	● Submissions due: Sunday, 11/21/2021 3:59 PM EST
Phase 3 Live Test	● DNS Submission open: Sunday, 11/21/2021 3:00 PM EST ● DNS Submission due: Sunday, 11/21/2021 3:59 PM EST
Phase 3 Report	● Due: Tuesday, 11/23/2021 11:59 PM EST

Live Test Schedule

Submit DNS for Live Test

Information

Time	Task	Description
4:00 pm	DNS Submission	Submit your DNS for the Live Test before 4pm
5:30 pm - 5:35 pm	DNS Validation	Validate your DNS. This is the last chance to update your DNS for the Live Test

Live Test

Information

Time	Value	Target	Weight
6:00 pm - 7:00 pm	Warm-up (M1 only)	0	0%
7:00 pm - 7:30 pm	M1	40000	15%
7:30 pm - 8:00 pm	M2	25000	25%
8:00 pm - 8:30 pm	M3	10000	25%
8:30 pm - 9:00 pm	Mixed (M1,M2,M3)	15000/10000/2000	5+5+5 = 15%

Team Project Phase 3

- Use only <u>AWS managed services</u>
- Web-tier
  - ECS with Fargate
  - EKS with Fargate
  - EKS with managed node group
- Storage-tier
  - Amazon RDS
  - Amazon DynamoDB
  - Amazon Elasticache
  - Others, check writeup for whitelist
- Teams should explore the managed services provided by AWS to come up with a solution

Task Reminder

- Throughput targets:
  - M1: 40,000 RPS
  - M2: 25,000 RPS
  - M3: 10,000 RPS
  - Mixed queries M1/M2/M3: 15,000/10,000/2,000 RPS
- Bonuses
  - Top 5 teams with highest performance/cost ratio
    → 4% Bonus
  - Raking 1st/2nd/3rd/4th/5th in the live test
    → 5%/4%/3%/2%/1% Bonus
  - Achieve target throughput for all tests
    → 2% Bonus

Total Budget Reminder

- Budget limit $120, double budget $150

	No Penalty	-10% Penalty	-100% Penalty
Total cost	< $120	$120 - $150	$150+
Development cost	< $90	$90 - $120	$120+
Live Test cost	~ $30	~ $30	~ $30

- Use GCP and Azure for ETL
- Use Spot instances wisely

Hourly Budget Reminder

- Your web service should not cost more than **$1.28/hour**

- This includes:

  - EKS / ECS cost

  - Database cost

  - ELB cost - excluding LCU-hour cost

  - See writeup for details

Resource Constraint Reminder

- No self-managed EC2 instances and EBS volumes during the Live Test!
- Rule of thumb: See resources in the EC2 dashboard besides managed node groups? NOT ALLOWED
  - https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#Instances
  - https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#Volumes
  - If you are installing MySQL as Kubernetes Deployment, StatefulSet, Helm chart → **STOP**
- T-family instances are not allowed, as their performance are not consistent

Hints for loading ETL results

- Perform ETL on Azure or GCP
  - Save results as TSV/CSV to cloud storage (e.g. Azure Blob Storage, GCP Cloud Storage)
  - Refer to [Spark docs](#)
- Launch your database
  - Use Terraform to reduce risk of human error (e.g. inconsistent configuration, forgot tagging…)
- Use a temporary EC2 instance to import
  - Use azcopy or gsutil to download ETL result
  - Make a backup on an S3 bucket
- Make a snapshot of the database

Hints

- Not all fully managed services for web-tier are suitable for Live Test. Some might not provide enough flexibility and result in higher total cost
- Consider the characteristics of the query before deciding on which managed database service to use
- High performance/cost ratio is preferred. Read the pricing of each service carefully and try your best to achieve the highest possible ratio
- Make sure you understand the concept of VPC and Security Group. Don't open your database to the world!

Best Wishes!!!