# DS561 Syllabus

Leonidas Kontothanassis

# Course Objective

Learn both the fundamentals and practical aspects of cloud computing (compute, storage, networking) in a real public cloud environment.  There are no required textbooks outside of the notes and publicly available papers and reference materials linked from these notes.

# Logistics

This syllabus: https://tinyurl.com/ds561-fall2023

Piazza: https://piazza.com/bu/fall2023/ds561, Password: nso2pbzspof

Gradescope: https://www.gradescope.com/courses/590893, Password: PWPZDZ

Google Platform Accounts: https://tinyurl.com/ds561-accounts

Me: kthanasi@bu.edu, CDS1641, Office hours: M/W 4:00-6:00pm @CDS1641

TAs: sahithi@bu.edu, vaishv@bu.edu Office hours: TBD

# Course policies

You are welcome to search the internet for help in all possible ways.  You must understand your solution and be able to explain it in your writeups for the assignments.

10 Homeworks (see list at the end of this document).

Grade will be determined 60% homeworks, 35% final exam, 5% attendance.

Code of conduct: https://www.bu.edu/academics/policies/academic-conduct-code/

Using AI bots: https://www.bu.edu/cds-faculty/culture-community/gaia-policy/

# Lecture List

Lecture 0: This document

Lecture 1: (General Cloud concepts)

Lecture 2: (GCP tools)

HW 1 handed out

Lecture 3: (File Systems)

Lecture 4: (Google Cloud Storage interfaces and APIs)

HW 2 handed out

Lecture 5: (Cloud Storage fundamentals)

Lecture 6: (Types of compute on the cloud)

# Lecture List

Lecture 7: (Messaging on the cloud)

HW 3 handed out

Lecture 8: Google Compute Engine APIs and options

Lecture 9: Operating System and Cloud Scheduling

Lecture 10: Virtual Machines

HW 4 handed out

Lecture 11: SQL basics, relational databases

# Lecture List

Lecture 12: Cloud SQL Google APIs tutorial

Lecture 13: Database Optimization

HW 5 handed out

Lecture 14: BigQuery Theory and Practice

Lecture 15: Spanner Theory and Practice

HW 6 handed out

# Lecture List

Lecture 16: Map Reduce

Lecture 17: Apache Beam, Cloud Data Flow tutorial

HW 7 handed out

Lecture 18:  IP, TCP, BGP

Lecture 19: DNS, SSL,HTTP

Lecture 20: Virtual Private Cloud and Load Balancers

Lecture 21: Virtualized Networking (Andromeda)

HW 8 handed out

# Lecture List

Lecture 22: Containers, Docker

Lecture 23: Docker and Kubernetes on GCP

HW 9 handed out

Lecture 24: Google Cloud Deployment Manager

Lecture 25: Markup languages, XML and YAML

HW 10 handed out

Lecture 26: Recap and summary

# Homeworks

Homework 1:
https://docs.google.com/document/d/1gcPin4X9M9Nj1fw5PBLqF47K_EETDjIWYrZj_SS69tk/edit

# Homeworks

Homework 2:
https://docs.google.com/document/d/1pco4Hj5DFViIskyh9Fhear7rLnzyzve4w58t-375-YA/edit

# Homeworks

Homework 3:
https://docs.google.com/document/d/1KMWRzsST_efD5hkBH4TPzKpgtOc1Oyk_cw9AAURwLCc/edit

# Homeworks

Homework 4:

Create a VM instance that can do the same as your cloud function.  This is much more complicated as it means that you have grab a VM, install the relevant software on it, set up DNS for it, write a web server, open firewall ports etc.  Run your client to issue a few hundred requests to demonstrate relevant serving and pub sub behaviors. Report your total spend on this homework.

# Homeworks

Homework 5:

Modify your webserver from HW4 to extract information from the requests (country, ip, age, gender, income, timeofday) and store them in a Cloud SQL database.  This is also quite complicated as you need to create the database, add the database interface into your webserver, install the SQL client on your VM and make sure permissions are set correctly.  Run your client to fetch 100K requests and store 100K records in the database. You should continue to publish to pub sub channels when requests arrive from "banned" countries.

# Homeworks

Homework 6:

Write a simple program that can retrieve the data from the database, and build 2 models that use some of the fields to predict some of the other fields.  One model should use client IP to predict the country, and another should use country, age, and gender to predict income.  You get to choose what kind of model you want to use.  You should aim for 99+% accuracy for the first model and 80+% accuracy for the second.

# Homeworks

Homework 7:

Use Apache Beam and Cloud Data Flow to do some data processing on the files in your cloud bucket.  In particular you should find and print out the top 5 files with the most incoming links. Bonus: If you can compute the pagerank score of pages using Cloud Data Flow (may be considered at the end of the class)

# Homeworks

Homework 8:

Modify homework 4 to create 2 servers in different zones and place them behind a load balancer (network or http options exist). Run your client for 10K requests against this setup.  Modify your servers to return their identity as a header and modify the client to count how many requests go to each server behind the load balancer. Show how the overall setup remains resilient when one of the servers behind the load balancers is taken out of service. Think of how you will establish health checks between the load balancer and your VM servers.

# Homeworks

Homework 9:

Migrate homework 4 to use Kubernetes containers instead of VMs.  You should drop the pub/sub functionality as setting up pub/sub permissions with GKE is quite complex.

# Homeworks

Homework 10:

Write a declarative configuration that can deploy the parts of Homework 5 (GCS buckets, VM web server, Cloud SQL database,PubSub) using Google Cloud Deploy manager