Project 2
Assessing hosting options
Due 10/16/17 11:59 pm CT

Your company, Aggie Startup Consultants, has a new prospective client: a startup in College Station that wants your help in implementing a new mobile app and its backend services. The company is in stealth mode, and it has been very successful in raising venture capital. They want to avoid information leaks about their new idea and its business model, so they are somewhat vague at this stage.

They want you to provide estimates on how much money they need to spend to host their application, even though they are revealing as little as possible about what kind of application they want to run. You need to write a report answering your prospective client's questions with as much precision as you can. If you need to make assumptions on aspects of the problem that are not specified, make sure that you state the assumptions clearly in your reports.

You can request clarifications questions on Piazza, but some aspects of the assignment are likely to continue to be quite vague. The issues that are indeed answered will appear in the Project 2 – FAQ (eCampus -> Content -> Project 2).

Your prospective client expects you to submit your project report by Monday, October 16th, 2017.

Background information

1. The terms "Backend" or "backend server" refer to the physical servers or virtual machines doing the "heavy lifting" for the application, i.e., running the services that implement the functionality. There is a web server running on the backend that receives requests through http connections and can invoke the appropriate applications and services also running on the backend.
2. Capital costs: costs to obtain equipment (including support plan). The purchases become part of the company's assets, and they are subject to a yearly depreciation rate.
3. Expense costs: payments for services or usage of products that do not become part of the company assets.

Provided information

1. The mobile app is able to upload selected videos and pictures to the backend, but only through access to the user's Google Photos, iCloud, Facebook, Instagram, YouTube, and Pinterest repositories of photos and videos, when granted permission by the user. This restriction means that your client can leverage the work done by other companies to flag inappropriate content.
2. The mobile app sends daily reports to the backend with the list of songs that the user listened to.
3. The app has a revolutionary way to compose thematic videos based on the user's library of photos/videos and her/his musical preferences. They claim that, when compared to their videos, the ones by Google Photos Assistant or Animoto look pathetic.
   Notice that Google Photos prepares videos without any request from the user. They automatically generate videos that they speculate will be considerate useful/interesting. Your client intends to do the same: continuous generation of new videos that are selectively presented to the users.
4. Even the simplest of the videos require computationally intensive image processing to identify the main "actors" or identify events or locations to serve as the theme in a produced video. For example, different movies are generated for each pet present in your pictures. Your clients created a new algorithm that extends the current deep learning techniques to achieve unprecedented precision in creating tags and meta-data to classify photos and videos.
5. The new image/video processing algorithm is currently optimized to run on either of two clusters available to them at Texas A&M:
   a. A cluster of four powerful machines: lots of memory, powerful CPUs and GPUs, access to SSD disks and a large network-attached storage appliance;
   b. A cluster of sixteen smaller machines: no GPU, no SSD, A smaller number of CPU cores, less memory.
6. The client does not want to specify what "lots of memory or "powerful CPU" means.
7. The client is not precise about how much benefit their algorithm gets from running on a more powerful hardware platform. They said that you should investigate costs for two scenarios: (1) the powerful cluster takes half the time to process images/videos and (2) the powerful cluster takes 25% less time than the simpler cluster to process images/videos.

8. Their code assumes Ubuntu 16.04 as the operating system. Most of their code is written in Python.
9. Texas A&M is investing in the company by providing the following deal for the first three years:
    a. allowing them to host as many computer racks as they may need at 5% of the average prevalent rates for San Francisco or New York;
    b. not charging for network traffic (either in, out, or intra-datacenter);
    c. offering disaster recovery for free, i.e., the datacenter will take nightly snapshots of the storage and application state for recovery if needed;
10. Initially, they will go to market only in Austin and College Station. The founders hope to have a curve of adoption similar to what Pinterest had in its first three years. Their assumptions:
    a. Through word-of-mouth, marketing, and university partnerships, your client estimates it will attract in the first month 5% of the students in the two prominent local universities.
    b. User profile:
        i. 10% of the user population are heavy content generators, uploading to their Google/iCloud/Pinterest/Instagram/Facebook accounts an average of 100 photos and 10 videos a day.
        ii. 70% of the user population upload an average of 5 photos and 0.3 videos a day.
        iii. 20% of the user population only consume videos, without uploading any photos or videos.
11. Users can make their videos publicly available. The founders expect to attract a few celebrities to their platform, so that their videos go to the top of trending lists, generating more usage of their app. This means that some of their videos will be extremely popular. Other generated videos will never be seen by anyone. To improve user experience, the access to the most popular videos should have as little latency as possible, while the access to unpopular videos can be less efficient.

Request for clarifications

You can post questions to your prospective client in our Piazza forum. The answers will be posted on Piazza and collected at the Project 2 – FAQ (eCampus, at the Content -> Project 2 section.)

What your client wants you to do

The client wants you to deliver two reports.

Report 1

A report with your recommendation (and its rational) on how they should approach their engineering platform for the first three years of operation:

- Should they do like Pinterest or Netflix, and host everything on a cloud platform such as Amazon AWS or Microsoft Azure or Rackspace? (Let's refer to this as cloud-based)
- Or should they do like Facebook, and have total ownership of their platform so that they are not constrained on how much performance they can get or what datacenter architecture is available? (Let's refer to this as TAMU-based)

Your report should show your estimates for capital and recurring expenses for equipment and engineering personal for both a cloud-based and a TAMU-based solution. It should also identify pros and cons for each of the approaches. Clearly state your assumptions and the load that your solution(s) would be able to handle.

At this point, your client is not looking for exact numbers. Instead, they want to see how agile your company is when faced with a challenge. For this exercise, they are giving you 8 hours to conclude this 1st report:

- the hours do not need to be contiguous;
- you have four business days to allocate the requested 8 hours to the project;
- the client expects you to conform to the Aggie's ethic;
- you are expected to use a timer to track the time you spend working on this problem;
- when the time tracked reaches the 8-hour limit, you stop working on the report;
- while not tracking the time, you can still think about the project, but you are not allowed to search for project-related data or do any back-of-envelope computations;
- you can declare yourself done with the task before the 8 hours if you want;
- your client will assess your report based on its process: what factors you are taking into consideration, how you obtain initial estimates for costs of hardware and services; what your assumption for the load to be handled by the system (number of users, number of pictures/videos to be processed, number of videos to be produced);
- You need to include in your report the time you spent working on this task.

You can choose any public cloud provider for data points in your analysis. If you do not have personal experience with any cloud provider, we recommend you use Amazon AWS (aws.amazon.com) for your analysis. Take into consideration not only the cost for renting servers but also the charges for storage and network.

Report 2

The client also wants your opinion on the ease-of-use of the Heroku cloud platform (heroku.com). You can deploy applications on Heroku for free (no credit card necessary for signing up.) They request that you go through the "Getting Started on Heroku" material (devcenter.heroku.com) and follow the instructions to deploy an app on Heroku's platform.

As we will learn later on the course, in IaaS (Infrastructure-as-a-Service) platforms such as Amazon AWS EC2, customers obtain a virtual server with the chosen characteristics, running the operating system of their choice. The customer can *ssh* into this virtual server, and it is up to you to decide which applications run on the server. In contrast, Heroku is a PaaS (Platform-as-a-Service) provider, and it offers customers the ability to deploy applications using one the runtime frameworks their support (e.g. Java, Python, Go, Node.js, Ruby, PHP). You can develop and test application functionality in your local development machines, and easily get new versions of the code deployed on Heroku's cloud.

Your task is to follow the "Getting started on Heroku" documentation to get an application deployed on Heroku. You can choose any of the languages available; your instructors did this exercise using Python.

Your report needs to include the following information:

- Time you spent on this task, i.e., going through all the steps described in the documentation, including installing software;
- Problems you faced and how you solved (or tried to solve) them;
- How useful this exercise is in terms of giving an idea of how Heroku's work;
- Your initial impressions on Heroku's pros and cons.

Project submission and rubric

You submit one pdf file containing your two reports. The submission "button" is located on eCampus -> Content -> Project 2.

Report 1 (70 points):

- Missing time tracking information: -70
- Level of analysis details
    - Insufficient detail: 0 points
    - Low: 30 points
    - Medium: 45 points
    - High: 60 points
- Quality of the writing (organization, clarity) – 10 points

Report 2 (30 points)

- Missing time tracking information: - 30
- Level of detail in the report
    - Insufficient detail: 0 points
    - Low: 10 points
    - Medium: 20 points
    - High: 30