# Assignment # 1: DBMS

# Find data storage, range and retrieval of ten famous websites.

## Youtube

### Storage:

1. As most people correctly wrote, Amazon *does not*use an RDBMS (a traditional relational DB, e.g. Oracle) to store product data. RDBMSs simply don’t support the required scale (amount of data and query throughput/latency).  
   According to external sites (e.g. this one) Amazon has on the order of half a *billion* products for sale, and that’s just the main US site (admittedly the biggest). According to this, it serves roughly 1B pages per day, which is roughly 10,000 pages per second on average, and much higher at peak.
2. As someone else observed, there is also no single DB used throughout Amazon. That was actually true in the very early days, when a single Oracle instance stored *everything:*product data, user accounts, orders, inventory… This hasn’t been the case for many, many years now.
3. However, there is a single “conceptual” DB that stores the vast majority of product information displayed on the site. It’s huge, super-fast, and extremely available.
4. This DB *isn’t*DynamoDB, or any other DB publicly available on AWS (RedShift, Aurora, etc.) It’s proprietary and private. This isn’t to say that you *couldn’t* build our catalog on top of one of those; I honestly don’t know.

### Range:

This is what Youtube says about how much content is being put up on Youtube:  
“60 hours of video are uploaded every minute, or one hour of video is uploaded to YouTube every second.”

OK, that’s a good start. And our reference video is 4 minutes.

youtube-dl -F (for listing all formats) shows me:

1. 37 : mp4 [1080x1920]
2. 22 : mp4 [720x1280]
3. 45 : webm [720x1280]
4. 35 : flv [480x854]
5. 44 : webm [480x854]
6. 34 : flv [360x640]
7. 18 : mp4 [360x640]
8. 43 : webm [360x640]
9. 5 : flv [240x400]

youtube-dl –all-formats (for downloading all formats) shows me that the total size of all formats/files is 340 MB of storage – for a 4 minute video.

I’m going to be a little optimistic and say that not all videos have a HD format and come in far fewer size options (in fact, probably a minority have HD video), so let’s say its about half that: 160 MB for a four-minute video (to make my calculation a bit easier to follow.)

Total storage for 1 minute = 160 MB / 4 minutes = 40 MB/min  
Total storage for 1 hour = 40 MB/min x 60 mins = 2400 MB = 2.4 GB

As per Youtube stats, 1 hour of video is being uploaded every second, so 2.4 GB of video is uploaded every second.  
\* 86400 seconds/day = 207.36 TB / day  
\* 365 days/year = 75.686 PB / year

So there you go. Approximately 76 PB of video data is stored in Youtube every year.

At the time of writing, [2 TB hard drives cost approximately $100 or less](http://www.amazon.com/Western-Digital-Caviar-Desktop-WD20EARS/dp/B002ZCXK0I).

So 76 PB x $50 / TB = $3,784,320 on storage costs  
= $3.8 million in storage costs

So there we go. Approximately **$4 million in storage costs per year**, using commodity hard drives. (I suspect the actual number could range widely, given the variable lengths of videos, formats, etc.)

Next, on to networking costs.

### NETWORKING

Again, from Youtube stats:  
“Over 4 billion videos are viewed a day”  
“Over 3 billion hours of video are watched each month on YouTube”

This is a bit more interesting – not all stored videos are served up, in fact, only one format of the video is selected and shown to the user.

Let me choose the [360×640] MPEG4 of the reference video, which is 40 MB in size. Also, let’s assume the user only watches 75% of the video or only that much is streamed over the network.

40 MB x 0.75 = 30 MB for 4 mins of video  
30 MB / 4 mins \* 60 min = 450 MB / hour

OK, over 3 billion hours of video are watched each month of Youtube. That is:  
100 million hours a day = 450MB / hour \* 100M = 45 PB/day  
Per month = 45 PB \* 30 = 1350 PB = 1.35 EB  
Per year = 1.35 EB \* 12 = 16.2 EB (exabytes!)

Wow. So 16.2 exabytes per year are sent through Youtube alone according to my calculations.

Now, I suspect I’m widely off mark here. For example, in [Wikipedia’s entry on Exabytes](http://en.wikipedia.org/wiki/Exabyte), it says:  
“Internet video will generate over 18 exabytes per month in 2013.”

And from [this PC Magazine report](http://www.pcmag.com/article2/0,2817,2395372,00.asp), it looks like Youtube only accounts for 10% of Internet traffic. Note that that’s all of Internet traffic, so video is probably a great majority of that (if “HTTP”, according to the report, is 20%, then video is at 80%).

Still, I appear to be off by a factor, let’s say of 4x, so let’s say Youtube is only pushing 5 EB a year.

Amazon’s Cloudfront CDN charges $0.12 per GB for low-end customers. I suspect that Google/Youtube, through their combination of in-POP networks, working with ISPs, etc, is able to keep that cost down. Let’s say $0.01 per GB.  
So $0.015 per GB = $15 per TB = $15,000 per PB = $15M per EB  
Cost = 6 EB \* $15M / PB = $90,000,000  
= $90 million

So there we go. Approximately**$90 million a year in networking costs**. (Again, this number could vary wildly given the video sizes, network bandwidth required and most importantly aggressive CDN networking/technology solutions that I’m sure Google is deploying, as well as low CDN costs and any existing deals Youtube has with ISPs to reduce traffic for themselves and ISPs.)

So there we go, combining the two costs:  
$4 million (storage) + $90 million (networking) =**~$100 million per year in networking+storage costs**

This does not include the server count/costs for that, or personnel.

And I have to clarify again, that this is a very rough estimate, and I wouldn’t be surprised my numbers were off by a margin of 0.5x to 4x

## Facebook

### Storage:

They do it with servers. A lot of them. Giant data storage facilities in Prineville, OR keep track of everything from the events you’re “attending”, all the way down to a status like (or love, haha, wow, sad, or angry face). In total, recent estimates put Facebook’s Oregon facility storage capacity at upwards of 300 petabytes, which would be enough to hold about 100 billion average-sized photos. The good news is this space is not going to run out anytime soon as the social media giant has both improved its storage efficiency and is nearing completion of its largest data center yet (at over 450,000 square feet).

What exactly is in these data centers and why should you care? First, if you were to take a walk through the building, you would be met with row after row of glowing blue towers. These towers each contain the hard drives that store and record all of your information. Sometimes that information becomes stale and hasn’t seen any action in a while (think: that music video you thought was so funny/cool in 2008, etc.). In these cases, it goes into cold storage until if/when it is needed again. This is important, mainly, because if Facebook didn’t have a place to put everyone’s information, then every time you logged off and logged back in, you’d have to start all over again. It would be impossible for Facebook to remember who you are. There’s a slot (or few million) reserved just for you to be able to share, comment, or post to your heart’s desire. And if you’re worried that anyone working for Facebook can easily access all your information, don’t be. In fact there are rooms in the facilities that are so secure that not even people who’ve worked there for years could access them if they wanted to, without the proper credentials.

### Range:

### Retrieval: