**Assumptions:**
Number of users: 2B
Number of devices per user: 2
Number of users opting for 2-step verification: 10% ~ 200M
Assume an attachment is on average = 1 MB

Questions:

*1. How much storage does gmail need per day to store emails?*

Let's say each email has 200 characters, on average. A user receives emails from useful connections, companies and spam.
Assume 20 spam emails, 20 marketing emails and 10 useful emails, per user per day.

Email data = Emails * Characters * Users = 50 * 200 * 2B
= 20 TB

Attachment data = number of emails with attachments * average attachment size
Attachment data = 5% of all emails * 1 MB
Attachment data = 5% * 50 * 2B * 1 MB = 5 PB

So total space requirement is Email data + Attachment data = 20TB + 5 PB per day.
This is a naively optimistic estimate, since we must account for redundancy (to improve performance and fault tolerance).

Estimated total space requirement = (20TB + 5 PB ) * 3 ~ **15 PB per day.**

However, we can optimise our storage by taking the hash of the contents of each email and storing only one copy. This will avoid repeated entries in the DB for the same email copy.

Assuming all marketing and spam mails have just one copy now, the number of emails per person per day goes to 1 + 1 + 10 ~ 15.

Hence, the total space required will reduce proportionately, from 50 to 15 per user.
That's 15 * 15 / 50 = **4.5 PB**

You can further reduce the size using encoding, but mention this with pros and cons. Old emails can be archived and compressed to save space. The newer/frequently accessed ones are better uncompressed and cached.

Assuming compression of 50%, we get 4.5 * 50% ~ **2.5 PB per day**.

*2. How much data does gmail need in total to store user profile information?*

Assuming 2B users, each having a name, DOB and user email address:
Each name is 15 characters on average. DOB: 8 characters. Email address: 20 characters.
That's (15 + 8 + 20) * 2B ~ 100 GB

Let's assume about 10% users have a profile picture. Assuming 100 KB per picture: 2B * 100KB * 10% = 20 TB

Taking redundancy into account, we need (20 TB + 100 GB) * 3 ~ **60 TB** space in total.

*3. How much processing power does the virus detector need to check all attachments?*

The virus detector needs to run through each attachment uploaded.
Total attachments = 5% of all emails * 1 MB average size
= (1 / 20) * (15 * 2B) * 1MB = *1.5 PB everyday*

Each attachment has to be scanned, then run through a virus detector. This will require static checks(string matches) and potentially running on a sandbox for tricky viruses. Assuming the time required to be 5 I/O reads, we have =

1.5 PB * (5 I/O Reads) = (1.5 * 10^6 MB) * (5 I/O reads)
Assume each read to take 20 ms per MB.

= (1.5 * 10^9 MB) * (5 * 0.02 seconds per MB)
= (1.5 * 10^9 MB) * (0.1 seconds per MB)
= 1.5 * 10^8 seconds
= 1.5 * 10^8 / (24*60*60) days
~ 1.5 * 10^6 / (25 * 36) days
= 1.5 * 10^6 / (100 * 9) days
= 1.5 * 10^4 / 9 days
~ 1.5 * 10^3
= 1500 days

To get 1500 days of work done in 1 day, we need 1500 virus detector processes running in parallel. Assuming we want servers running at 50% capacity and possible spikes in load, we can provision 1500 * 4 = **6000 processes.**

*4. Similarly, how much processing power does the spam detector need?*

Total emails = 15 * 2B = 30 billion.
Size of each email = 200 bytes.
Total email data = 30 billion * 200 bytes = 6 TB.

Spam detection needs a classifier (example: Bayesian) or a neural network to identify spam. This takes time to run. Assuming time of 5 I/O reads:

= 6TB * (20ms per MB * 5)
= 6 * 10^6MB * 0.02 seconds per MB * 5
= 6 * 10^6 MB * 0.1 seconds per MB
= 6 * 10^5 seconds
= 6 * 10^5 / (24 * 60 * 60) days
= 6 * 10^3 / (24 * 6 * 6) days
~ 6 * 10^3 / (25 * 36) days
= 6 * 10^3 / (100 * 9) days
= 6 * 10^1 / 9 days
= 60 / 9 days
~ 6 days

We need 6 days of work to be done everyday. This requires 6 processes to be run in parallel.
Considering 50% capacity + spikey load, we need 6 * 4 = **24** processes in total.

We assume the emails being classified as spam or not spam by a local service. This service will process emails in batch/stream, and persist the results to multiple geographical data centers.

*5. How much contact data should be kept in cache?*

Assume 1% of the users to be active at any point in time. We have 2B /100 = 20 million active users.

Each has a contact list. Assume the 'hot' contacts to be the top 10. However, a lot of these hot contacts will be common amongst users. Instead of a fan out, we assume the top 10 to be always from the active users.

These active users will probably be repeated in each other's contact list. If they appear in 10 contact lists, we have 20 million active users / 10 = 2 million unique contacts to be cached.

2 M unique users * profile picture data, if cached = 2M * 100 KB = 200GB.

If we take 64GB machines, we need at least 200GB/64GB ~ 4 machines.

Taking fault tolerance = 3 and localized processing = 10, we get
= 4 * 3 * 10
= **120 machines** of 64 GB each.