#ASLI ENGINEERING

How Instagram efficiently serves HashTags?

SWIPE

BY

ARPIT BHAYANI

Serving HashTags ordered by count?

Instagram uses Postgres as its main DB.

For each HashTag it stores the count of "media" associated with it

hashtag	media_count
☰	☰

snow|

snowman	2m
snowwhite	1m
snowy	1m
snow	500k

* Instagram uses ElasticSearch for advanced usage and proposes to use Postgres for this one

Query:

```
SELECT * FROM hashtags
  WHERE name LIKE 'snow%'
  ORDER BY media_count DESC
  LIMIT 10;
```

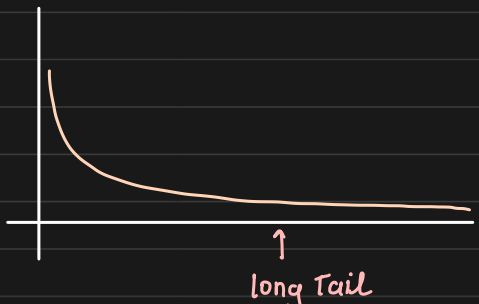
Executing this query have to go through a large number of rows, filter out rows, and then sort!

This query had to sort ~ 15000 rows

at scale this becomes a problem

Analyze

HashTags exhibit a long-tail pattern
i.e. fewer tags have large # photos
and large number of tags will have
very few photos



So, instead of going through all the hashtags,
can we only go through hashtags that are "popular enough"?

↳ Can we only index hashtags that has more than 100 photos?

Partial Indexes

Postgres supports partial indexes
i.e. keeping only a subset of data
in the index

```
CREATE INDEX CONCURRENTLY
on hashtags (name h_pat)
WHERE media_count > 100;
```

This substantially reduces the rows indexed
and hence query runs superfast!

With this change Postgre had to sort only
169 rows instead of 15000!!