Google Typeahead

Agenda

1. How to approach any HLD Problem in an interview
2. Design Google Typeahead.
3. Deep Dive (Problem Statements / Followups) related to Typeahead.
   - How to handle large reads & writes at same time
   - How to handle recency for suggestions

How to approach any HLD Problem in an interview (45 Mins)

0. Overview (Optional: only if you haven't heard about the interview wants you to design)

Min^m viable product

1. Functional Req → MVP features that should be supported

by the design you will come with

(Core features)

(4-5 core features)

2.) NON FUNCTIONAL REQUIREMENTS

(how those features should work)

for every feature in FR

⇒ Consistency V/S Availability

⇒ Consistency V/S Low-latency

3.) BoE Calculations (Back of Envelope)

[Estimation of Scale]

Approx

[Metrics]

1.) QPS / RPS : Query / Sec

Req / Sec

2.) [Storage]

50 GB

5 PB

RDB

Any other

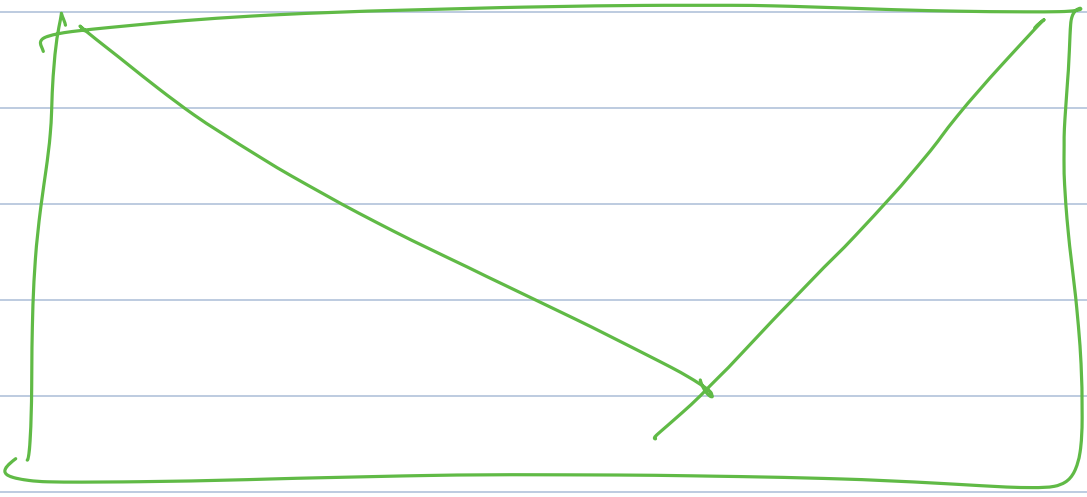
[Do we need Sharding ?]

files | images | video

3. N/W bandwidth (optional)

(5) uploading fees

DB ⇒ 1TB can be in 1 Machine



10000 users / day

$$\frac{10000}{24 * 60 * 60}$$

$$\rightarrow \frac{10000}{86400} \approx 100000$$

# of sec in a day = 10^5

# of days in an year = 400

$10^9 \Rightarrow$ 1 GB

$10^6 \Rightarrow$ 1 MB

$10^3 \Rightarrow$ 1 KB

$10^{12} \Rightarrow$ 1 TB

why ?

→ Sharding

- Is my sys (read heavy / write heavy / both)

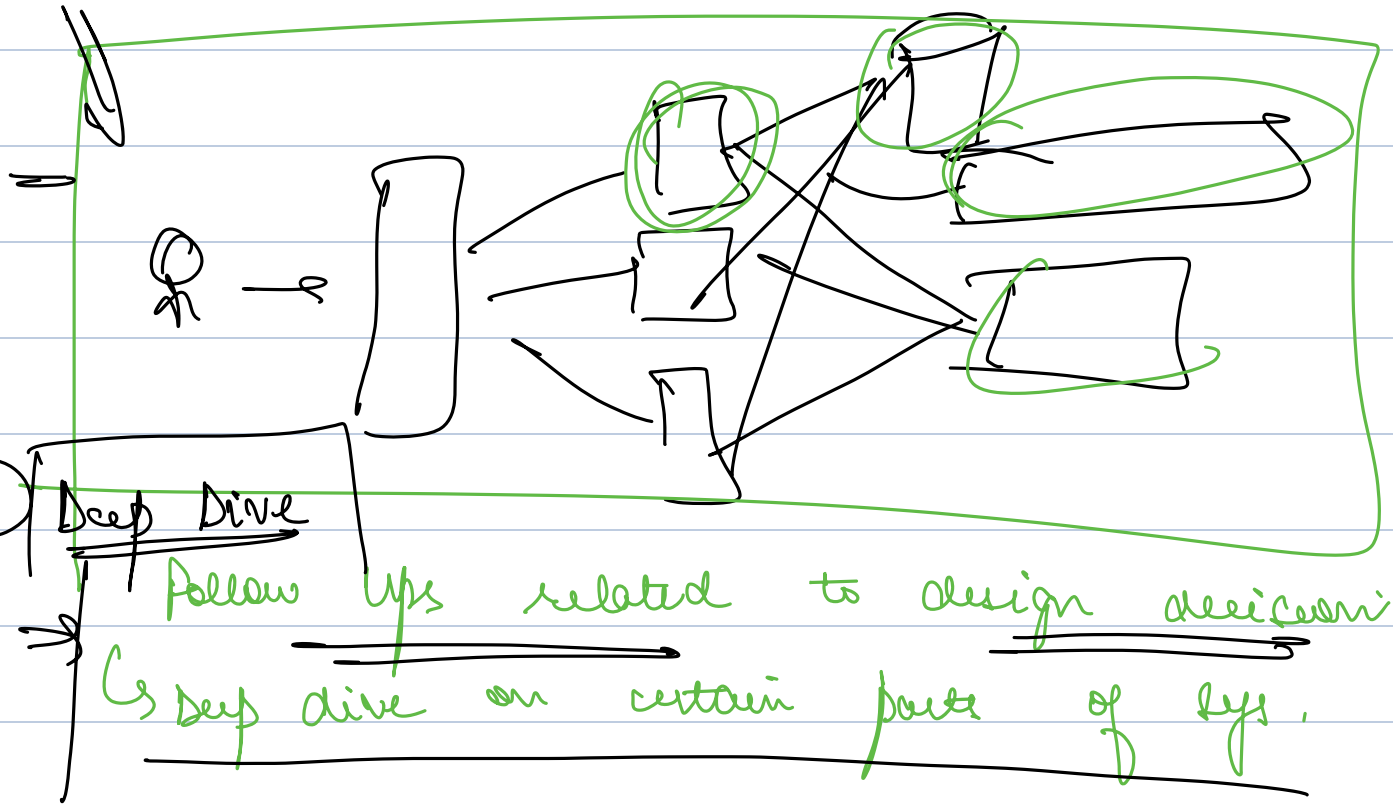
4.) API Design

→ first down the APIs that your sys has to support

get Suggestions( )

5.) Design

↳ block diagrams
↳ etc



6) Deep Dive

Follow ups related to design decisions
↳ Deep dive on certain parts of sys.

0.) Overview
1.) f^{n} Req
2.) Non f^{n} Req
3.) BoE Calculations
4.) API Design
5.) HLD
6.) Deep Dive

⇓

40 Mins

Design Google Typeahead

Overview ✓

wea
weat
tweality

F^n Requirements

1.) given a string get $sugg^n$ based on that

2.) Suggestions based on prefix

~~3.) spell check~~

~~4.) personalization~~

5.) 10 suggestions every time ← weather

6.) $sugg^n$ should be the top frequent queries with the prefix above

~~7.) recency factor~~

sac

sac

- hui conductor
- K
- ufic
- hui Banana
- ed Games

10 most occurring ones

Sac

hin tendulka

Sac

Donald Trump

till 2016

donald duck

1995

2017

donald

F^n Req

4-5 Mins

1. get Suggestions for Prefix
2. to Sugg^n
3. Recency Factor

Non F^n Req

Con 4/s

Cons 4/s

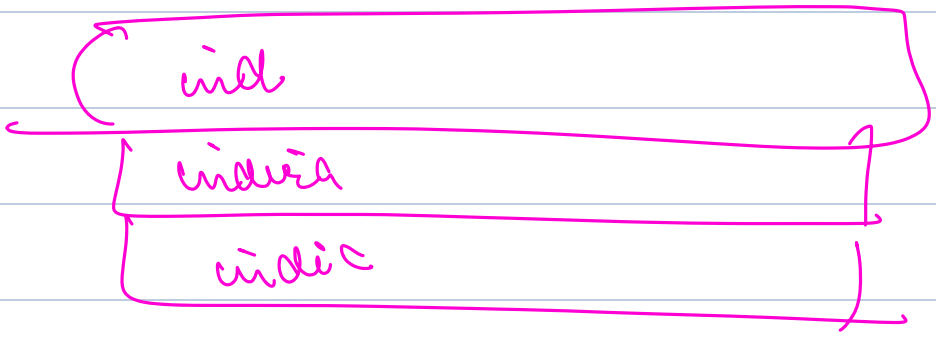
What core means?

ind

india 1031

indian 1032

indian



---

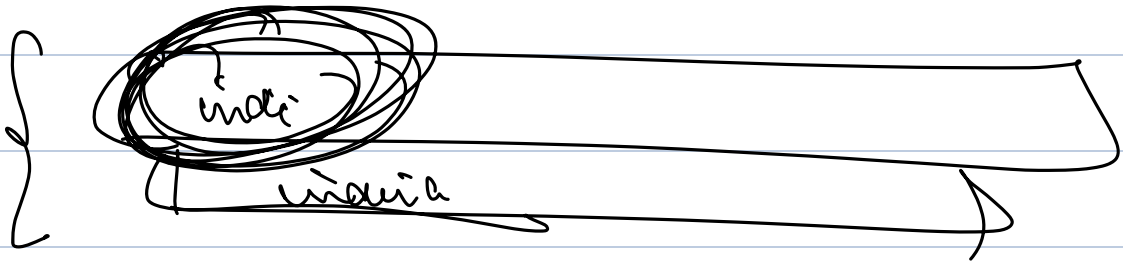
(3) BOE Calculations

"Guesstimate"

DAU = 2B

Or #searches / person / day ⇒ 10

= # searches ⇒ 20B



avg # of diff types of search queries per user = 6

→ # of search queries / day

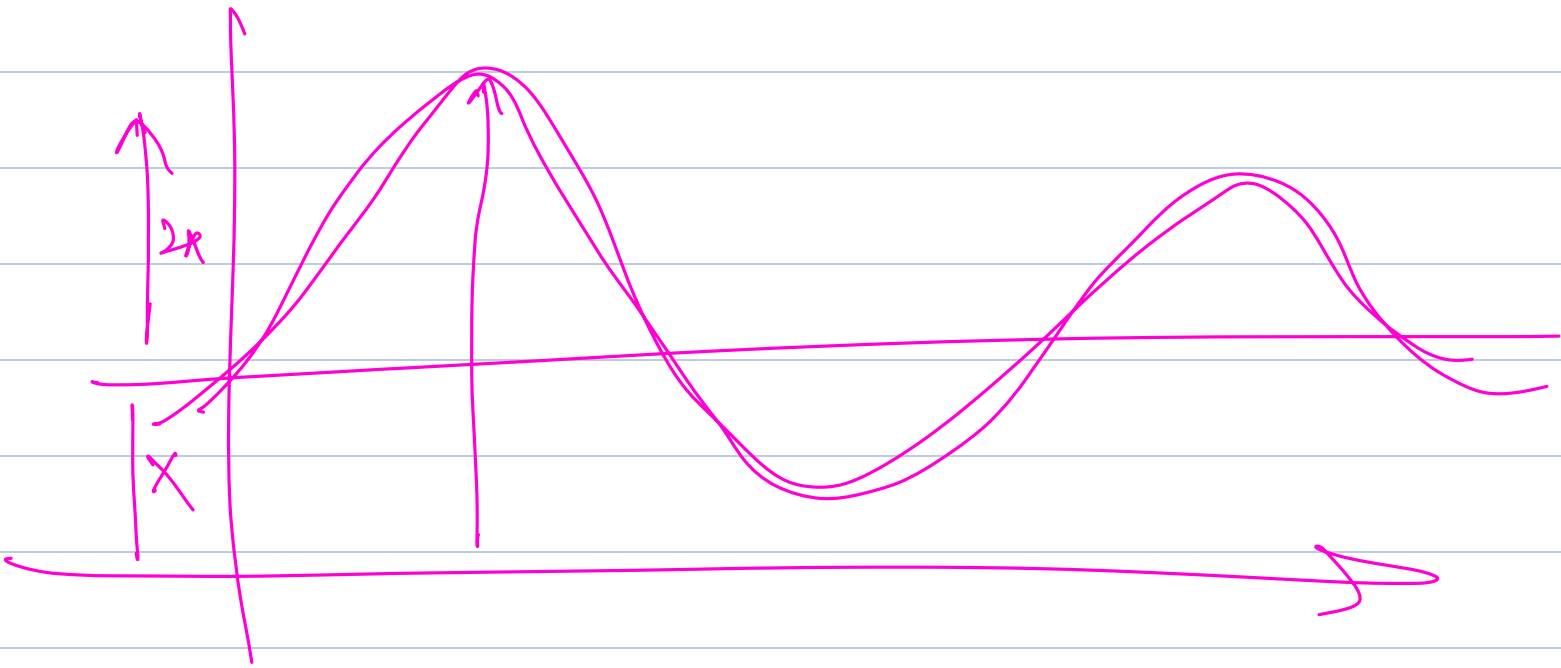
20B * 6 = 120B

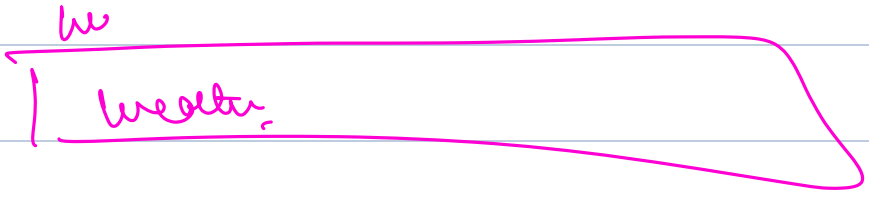
# of search queries / sec → $\frac{120 \times 10^9}{10^5}$

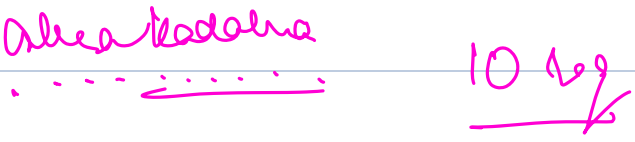
→ 1200000

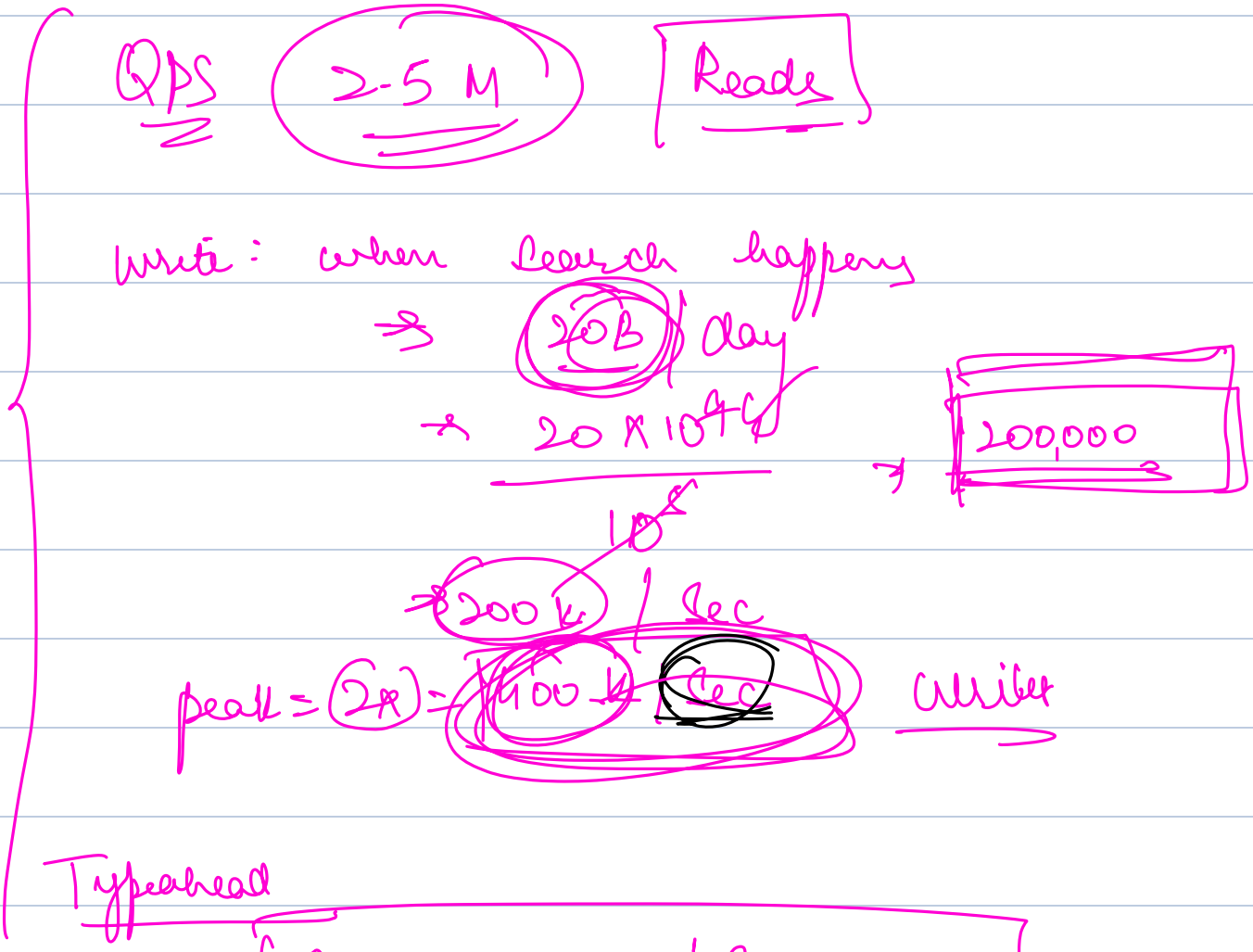
→ 1.2M / sec

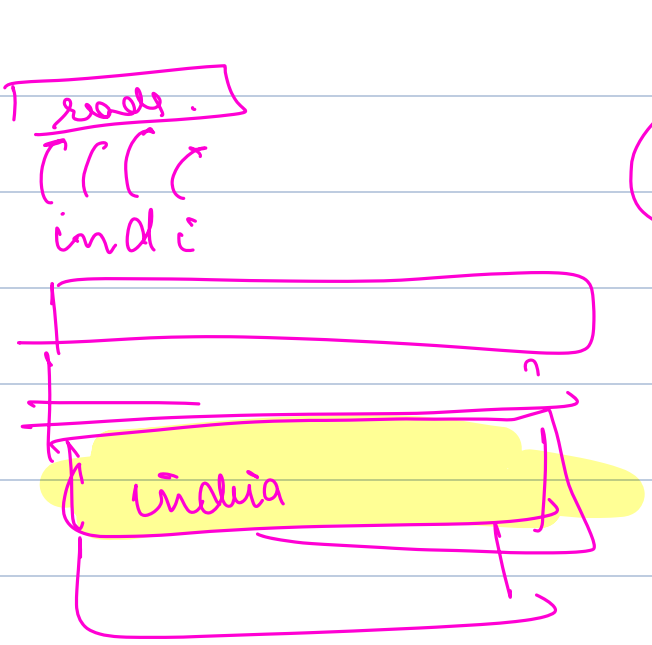
Assume peak = 2x average → 2.4 M / sec

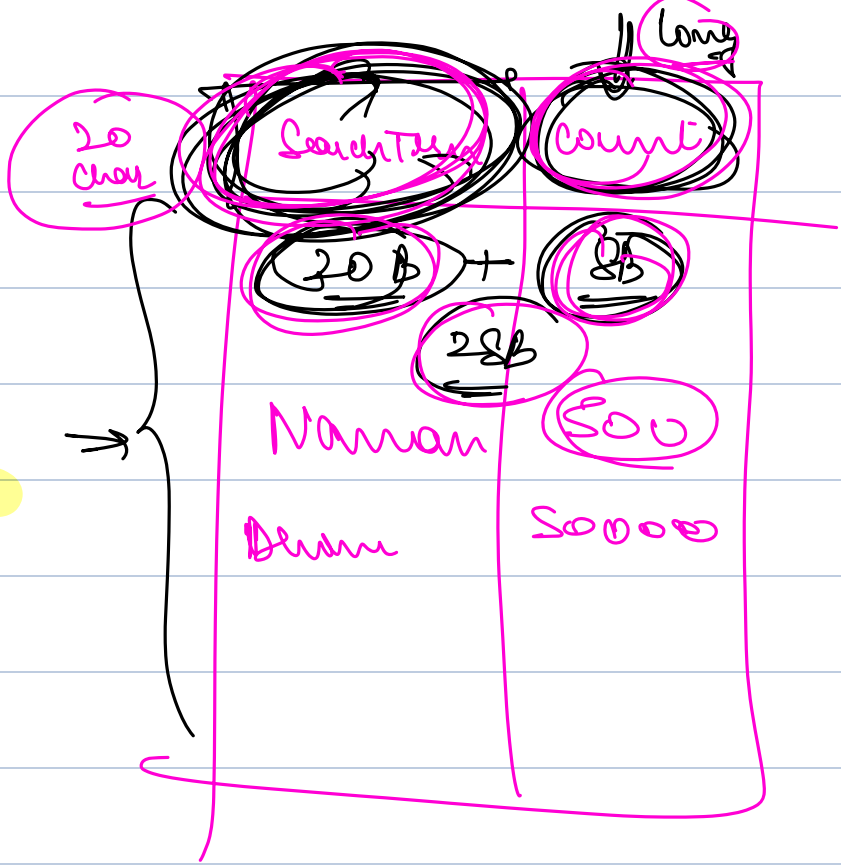

2R
X


w
Write.


Read/Write
10 TB


QPS
2-5 M
Reads
Write: when search happens
20B day
20 X 10^9
200,000
10
200K / sec
peak = (2x) = 400K sec
Write
Typeahead
Read → 2·5 M / Sec
Writes → 500K / sec





Storage

⇒ 28B Search Term

1 day ⇒ 20B

⇒ 10% unique ⇒ 2B terms / day

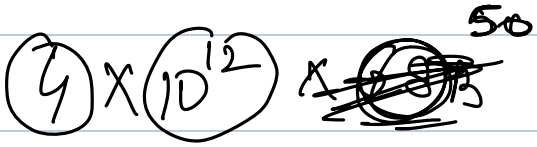
we should plan for 5 years

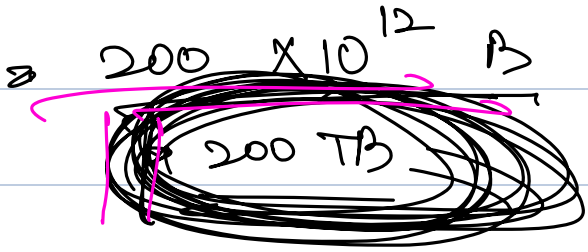
$2B \times 5 \times 400$

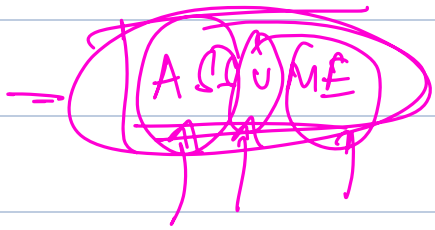
$\rightarrow \frac{2B}{2000}$

$\rightarrow 2 \times 10^9 \times 2 \times 10^3 \rightarrow 4 \times 10^{12}$

4T each term in 5 years







⇒ Sharding

↳ more than 1 TB

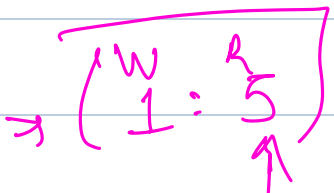
BoE Calc

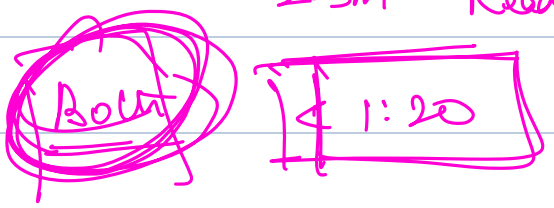
Sharding ↳

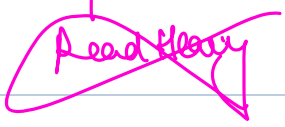
Read v/s Write

500K Write

2.5M Reads







5 : 100