

EP27: Stack Overflow Architecture. Also...



THERESA

OCT 8, 2022



145



5



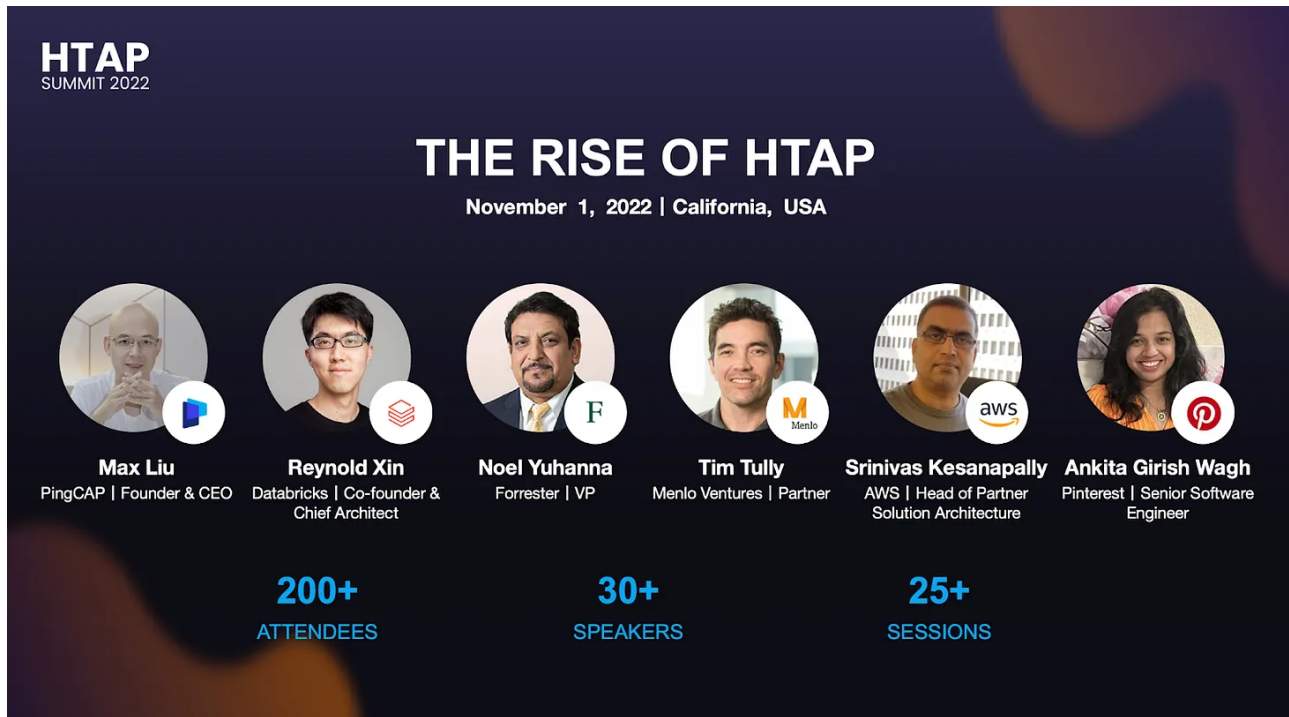
Share



This week's system design refresher:

- Stack overflow architecture
- iQIYI database selection trees
- Latency Numbers Every Programmer Should Know for the 2020s
- Row-based DB vs. Column-based DB

HTAP Summit 2022 is coming soon! (Sponsored)

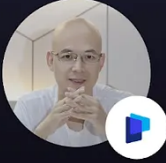


The banner for HTAP Summit 2022 features a dark blue background with a subtle pattern of light blue dots. At the top left is the HTAP SUMMIT 2022 logo. The main title 'THE RISE OF HTAP' is centered in large white letters, with the date and location 'November 1, 2022 | California, USA' below it. A row of six circular headshots of speakers is displayed, each with a small logo to its right. Below the headshots are the names and titles of the speakers. At the bottom, three statistics are listed: 200+ ATTENDEES, 30+ SPEAKERS, and 25+ SESSIONS.


HTAP SUMMIT 2022

THE RISE OF HTAP


November 1, 2022 | California, USA




Max Liu
PingCAP | Founder & CEO




Reynold Xin
Databricks | Co-founder & Chief Architect




Noel Yuhanna
Forrester | VP



Tim Tully
Menlo Ventures | Partner



Srinivas Kesanapally
AWS | Head of Partner Solution Architecture



Ankita Girish Wagh
Pinterest | Senior Software Engineer

200+
ATTENDEES

30+
SPEAKERS

25+
SESSIONS

We're talking about **HTAP Summit 2022**, the very first in-person conference on Hybrid Transactional / Analytical Processing. This promises to be a disruptive technology in the database world. So, dive in and discover more about this emerging tech!

Hear from 30+ database industry leaders and developers from companies and universities, such as Amazon, Databricks, Forrester, Block, Pinterest, PingCAP, Vercel, UW-Madison, UC-Berkeley, and many more.

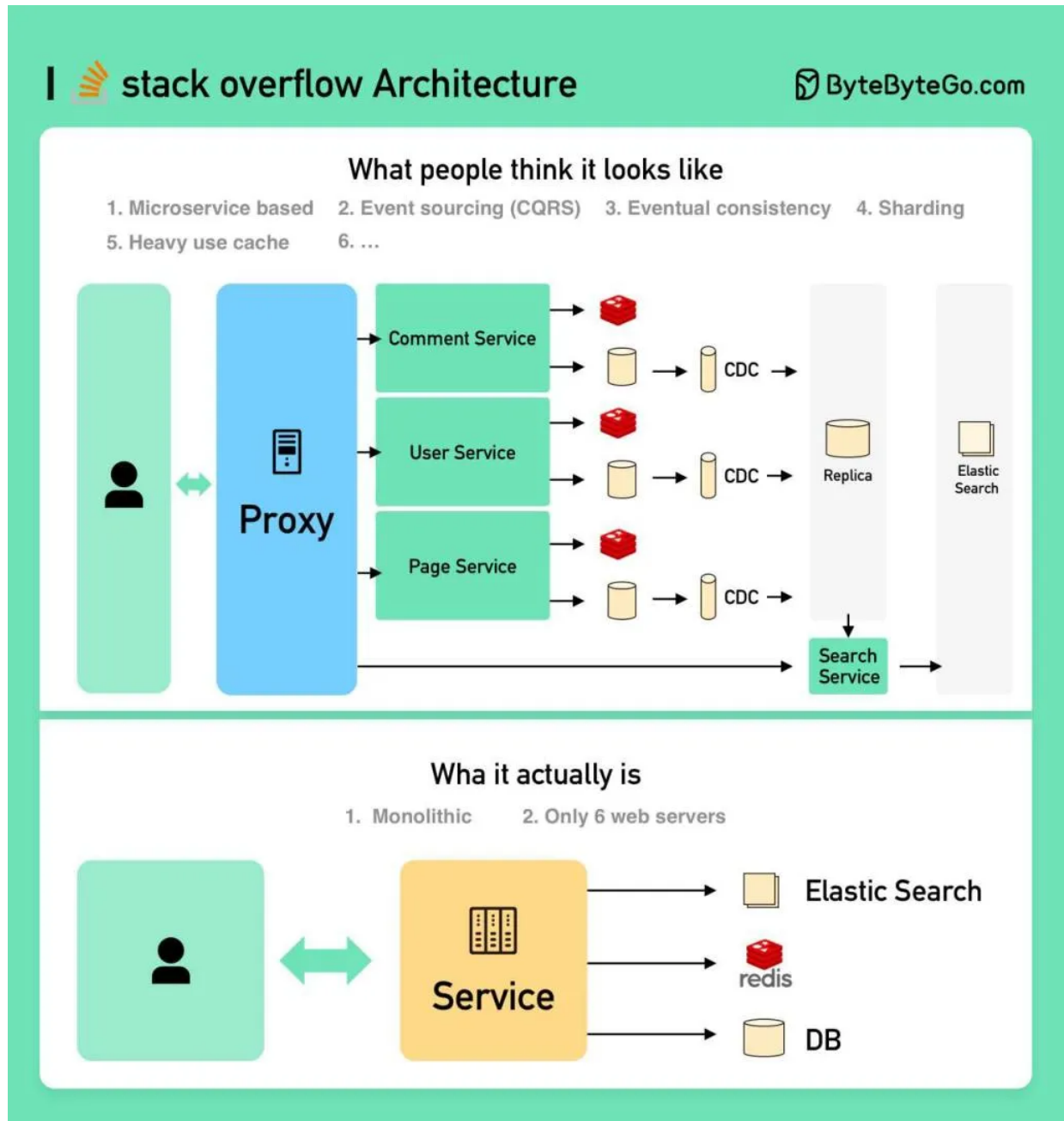
The Best part? It's **FREE**.

Date: **November 1** at the Computer History Museum, Mountain View, Bay Area, CA.

HTAP Summit 2022 organized by PingCAP features 30+ content-rich sessions on HTAP databases, including core infrastructure technologies, use cases, best practices, ecosystem, hands-on workshops, and keynotes.

How will you design the Stack Overflow website?

If your answer is on-premise servers and monolith, you would likely fail the interview, but that's how it is built in reality!



What people think it should look like

The interviewer is probably expecting something on the left side.

1. Microservice is used to decompose the system into small components.
2. Each service has its own database. Use cache heavily.
3. The service is sharded.
4. The services talk to each other asynchronously through message queues.
5. The service is implemented using Event Sourcing with CQRS.
6. Showing off knowledge in distributed systems such as eventual consistency, CAP theorem, etc.

What it actually is

Stack Overflow serves all the traffic with only 9 on-premise web servers, and it's on monolith! It has its own servers and does not run on the cloud.

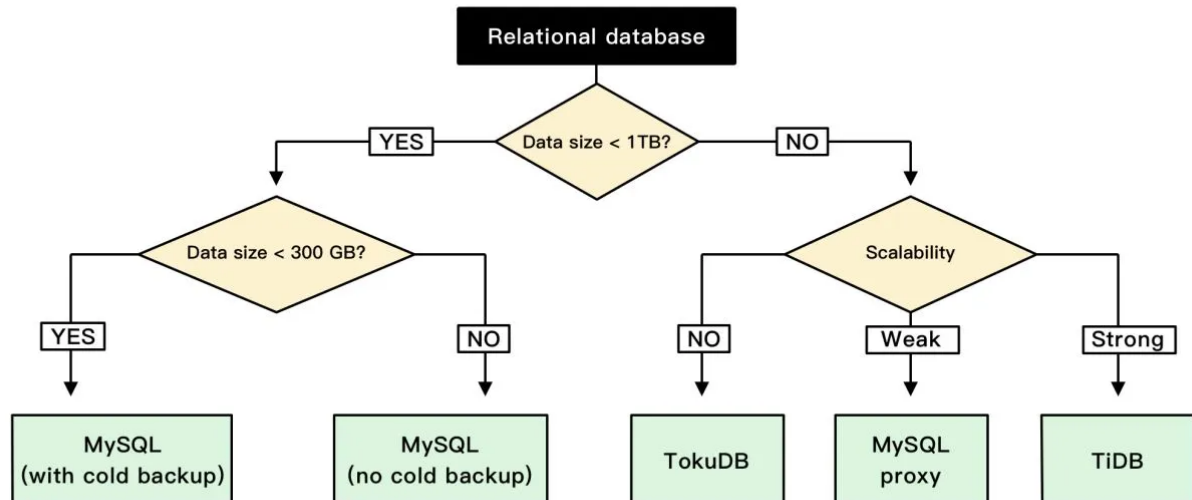
This is contrary to all our popular beliefs these days.

iQIYI database selection trees

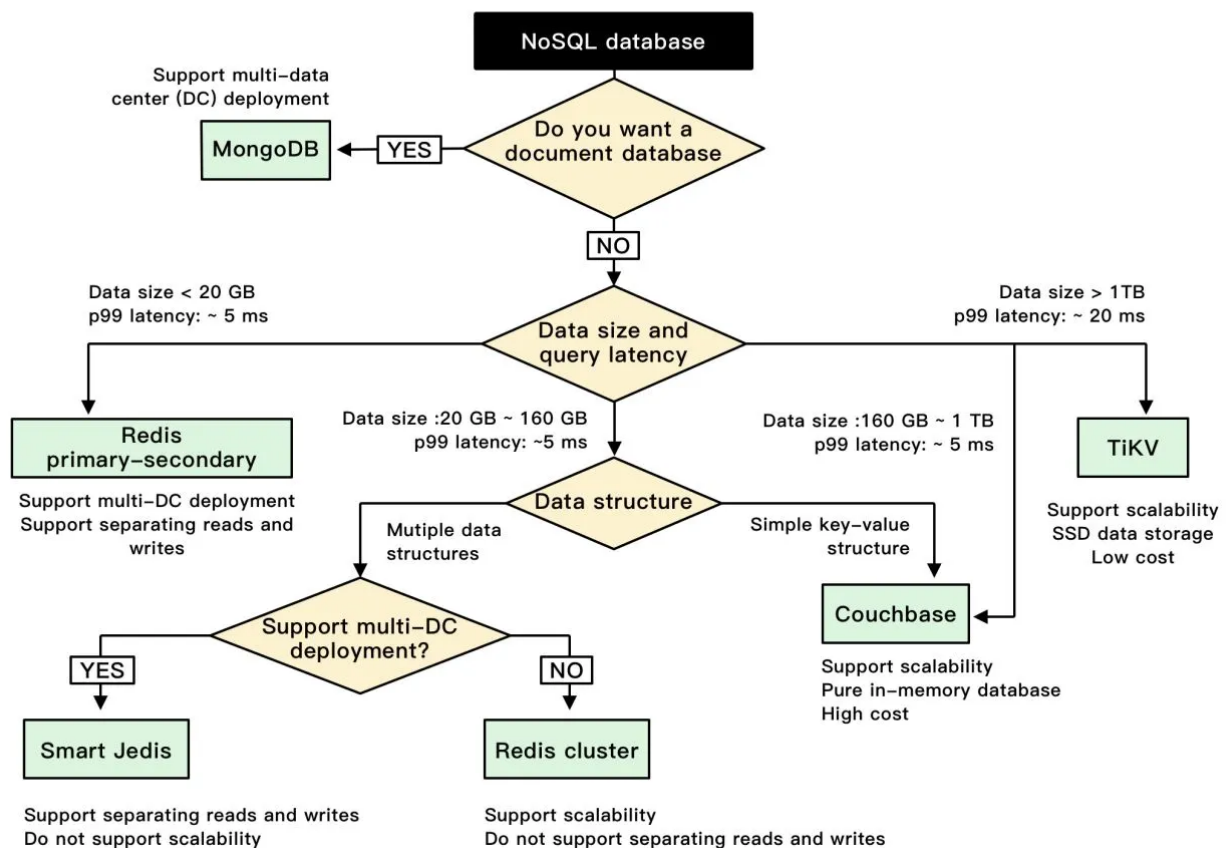
One picture is worth a thousand words.

iQIYI is one of the largest online video sites in the world, with over 500 million monthly active users. Let's look at how they choose relational and NoSQL databases.

Efficiently choosing a relational database



Efficiently choosing a NoSQL database



The following databases are used at iQIYI:

- MySQL
- Redis

- TiDB: a hybrid transactional/analytical processing (HTAP) distributed database
- Couchbase: distributed multi-model NoSQL document-oriented database
- TokuDB: open-source storage engine for MySQL and MariaDB.
- Big data analytical systems, like Hive and Impala
- Other databases, like MongoDB, HiGraph, and TiKV

The database selection trees below explain how they choose a database.

Latency Numbers Every Programmer Should Know for the 2020s

This concept was originally presented by Jeff Dean. We updated some of these numbers to more closely reflect reality in the 2020s. Absolute accuracy is not the goal. Developing an intuition of the relative differences is.

Latency Numbers Programmer Should Know: Crash Course System D...

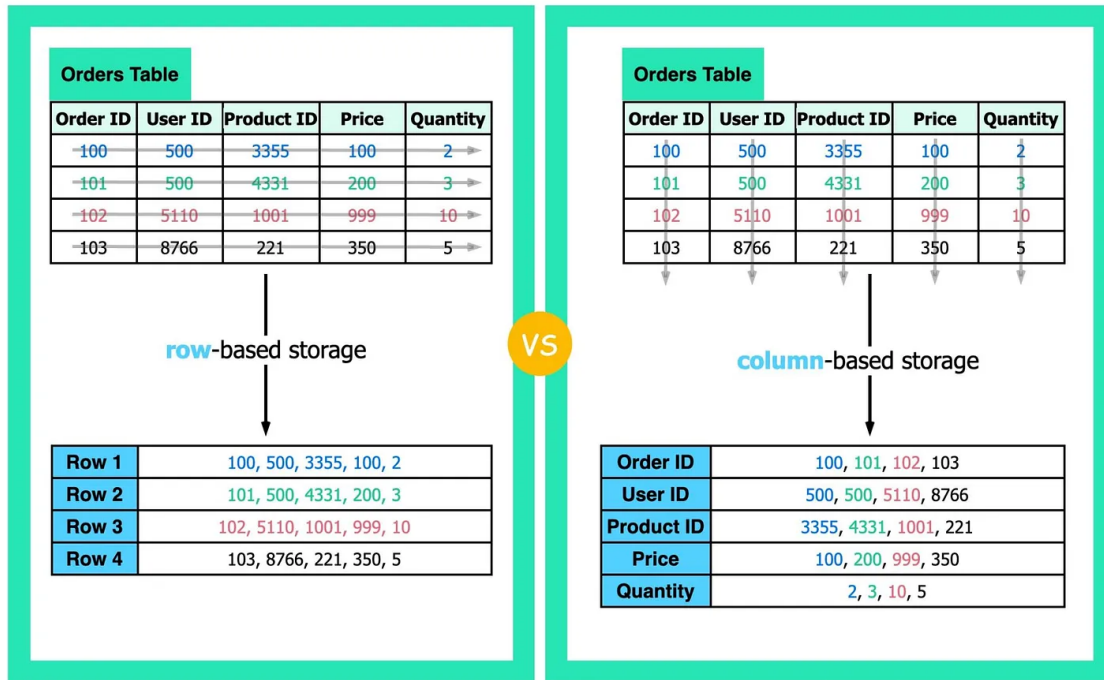


Why do we use column-based DB? Does column-based DB provide better performance?

The diagram below shows how data is stored in column-based DB.

Row-based DB v.s Column-based DB

 blog.bytebytego.com



When to use

1. The table is a wide table with many columns.
2. The queries and calculations are on a small number of columns.
3. A lot of the columns contain a few distinct values.

Benefits of column-based DB

1. Higher data compression rates.
2. Higher performance on OLAP functions.
3. No need for additional indexes

Got behavioral interviews? (Sponsored)



"Tell me about a time when..." Sometimes, the toughest interview questions aren't the technical ones. For behavioral interviews, RocketBlocks is here to help. Trusted by leading institutions like Stanford GSB and MIT Sloan.



145 Likes

5 Comments



Write a comment...



Kai Oct 10, 2022

Would like to see a further breakdown of the Stack Overflow system. I feel it may be oversimplified in the diagram, for instance, there must be a reverse proxy to distribute traffic between the different web servers?

♡ LIKE (4) 💬 REPLY ↗ SHARE



Julia Tong Mar 21

Can we have a further detailed evaluation of pros and cons of wide column DB, and which use case is the best to use them? Thanks

 LIKE  REPLY  SHARE

...

3 more comments...

© 2023 ByteByteGo · [Privacy](#) · [Terms](#) · [Collection notice](#)
[Substack](#) is the home for great writing