System Design (/courses/system-design) / Storage Scalability (/courses/system-design/topics/storage-scalability/) / Highly Available Database

Highly Available Database

Design a distributed key value store which is highly available and is network partition tolerant

Features:

> *“ This is the first part of any system design interview, coming up with the features which the system should support. As an interviewee, you should try to list down all the features you can think of which our system should support. Try to spend around 2 minutes for this section in the interview. You can use the notes section alongside to remember what you wrote. ”*

Q: What is the amount of data that we need to store?
A: Let's assume a few 100 TB.

Q: Do we need to support updates?
A: Yes.

Q: Can the size of the value for a key increase with updates?
A: Yes. In other words, its possible a sequence of keys could co-exist on one server previously, but with time, they grew to a size where all of them don't fit on a single machine.

Q: Can a value be so big that it does not fit on a single machine?
A: No. Let's assume that there is an upper cap of 1GB to the size of the value.

Q: What would the estimated QPS be for this DB?
A: Let's assume around 100k.

Estimation:

> *“ This is usually the second part of a design interview, coming up with the estimated numbers of how scalable our system should be. Important parameters to remember for this section is the number of queries per second and the data which the system will be required to handle.*
> *Try to spend around 5 minutes for this section in the interview. ”*

Total storage size : 100 TB as estimated earlier
Total estimated QPS : Around 100k

Q: What is the minimum number of machines required to store the data?

A: Assuming a machine has 10TB of hard disk, we would need minimum of 100TB / 10 TB = 10 machines to store the said data. Do note that this is bare minimum. The actual number might be higher if we decide to have replication or more machines incase we need more shards to lower the QPS load on every shard.

Sign up to view the complete problem.

Discussion

↑ cesar_augusto_guzman_alvarez (/profile/cesar_augusto_guzman_alvarez) 3 months ago
0 nice course. I only think it is a little bit long.
↓ reply

↑ jeffery (/profile/jeffery) 3 months ago
0 100T +
↓ reply

> ↑ jeffery (/profile/jeffery) 3 months ago
> 0 Sorry mistype, I wanted to leave this on the note area.
> ↓ reply

↑ henry_henry (/profile/henry_henry) 3 months ago
1 How exactly do peer-to-peer systems work? Doesn't there still have to be a high-level router that routes traffic to one of
↓ the systems, which represents a SPOF? Any good reads on these?
reply

> ↑ jax.teller (/profile/jax.teller) 3 months ago
> 1 The major idea behind a peer to peer system is that all the nodes are similar in nature/power. Here we use the idea in the sense that multiple nodes can be the owner of a data D. Depending on the settings, more than one node have to go down for this data to become unavailable. The high-level router you talk about can consist of multiple machines since it doesn't require any actual information about that data other than where it might be stored.
> ↓ reply
>
> > ↑ henry_henry (/profile/henry_henry) 3 months ago
> > 0 I see, I was originally a bit confused about consistent hashing using peer-peer systems. Out of curiosity, how do we get rid of "Single point of failure"? Will there not always be 1 SPOF regardless of what system we create? In a general sense, if a system goes down on any level of an architecture, doesn't there have to be a router of some sort that directs traffic?
> > ↓ reply
> > reply
> >
> > > ↑ kuzmin (/profile/kuzmin) about 2 months ago
> > > 0 You could have DNS-level routing that points to multiple load balancers
> > > ↓ reply

↑ henry_henry (/profile/henry_henry) 3 months ago
0 I see, I was originally a bit confused about consistent hashing using peer-peer systems. Out of curiosity, how do we get
↓ rid of "Single point of failure"? Will there not always be 1 SPOF regardless of what system we create? In a general sense, if a system goes down on any level of an architecture, doesn't there have to be a router of some sort that directs traffic?
reply

> ↑ henry_henry (/profile/henry_henry) 3 months ago
> 0 Meant to write this as a reply
> ↓ reply
>
> > ↑ munir_mehta (/profile/munir_mehta) 2 months ago
> > 0 I am assuming you are talking about load balancer who directs traffic in peer to peer system and
> > ↓ worried about that going down. Highly available system has redundant load balancer as well which will come back online if it doesnt hear back from master. Highly available system has everything redudant starting from load balancer to hubs and switches as well.

reply

↑ sarang (/profile/sarang) 3 months ago
0 For a read to be consistent(return the latest write), we need to keep W + R > P.
↓ reply

↑ sarang (/profile/sarang) 3 months ago
0 first thing first, what is P? is it number of shards for a key?
↓ reply

↑ rajputnr (/profile/rajputnr) 3 months ago
0 P is the replication factor here as explained above
↓ reply

↑ sumit_007 (/profile/sumit_007) 3 months ago
0 Yes, It is the number of shards which contains data for a paticular key.
↓ reply

↑ kartikeya_singh (/profile/kartikeya_singh) about 1 month ago
0 P is the no. of peers(machines) in a shard, shards don't inter-replicate data.
↓ reply

↑ rajputnr (/profile/rajputnr) 3 months ago
0 P is the replication factor here as explained above
↓ reply

↑ sumit_007 (/profile/sumit_007) 3 months ago
0 Does P, W and R predefined ? Or they change dynamically as per new machines addition/deletion ?
↓ Ex: For a total of 15 machines, I keep 'P' to be 5, 'W' and 'R' both to be 3.
Q1. What is the criteria to decide 'P' as 5 out of total 15 machines ? Any good articles/paper ?
Q2. How to handle if machines left available goes down below 'W' ? How sacrificing data consistency means here?
Q3. After addition of 10 more machines, will value of P, W, and R remains same ?
reply

↑ jax.teller (/profile/jax.teller) 2 months ago
0 Q. Does P, W and R predefined ? Or they change dynamically as per new machines addition/deletion ?
↓ A. P, W and R should be defined according to the use case for the system. For example, a system with high reads and low writes, we can keep R as 1 and W as P(hypothetically) so that we spend less time on reads while it is okay to spend more time on writes. P, W and R should be adjusted so that it represents how important the data is(P) and what is the ratio of reads vs writes.

Q. How to handle if machines left available goes down below 'W' ? What sacrificing data consistency means here?
A. Sacrificing data consistency would mean that if we are only able to perform a write on a subset of W nodes, while reading we may get an outdated value reading from R nodes. For example, assuming P = 5, W = R = 3. Let's name these machines M1, M2 .. M5. Assume before writing a value machines M1, M2 and M3 dies. For this write, we will have to do with M4 and M5. Now after some time, M1, M2 and M3 comes up. Now, if we try to read the earlier value and if we only read from M1, M2 and M3, we won't have the latest value for the data and hence we are compromising consistency.

Q. After addition of 10 more machines, will value of P, W, and R remains same ?
A. Yes. P, W and R are parameters about a single data value itself. More machines means more data to store with the same configuration.
reply

↑ kaidul (/profile/kaidul) about 2 months ago
0 Size of the value for a key can increase as mentioned earlier. But the solution wasn't discussed :O
↓ reply

↑ tersduz88 (/profile/tersduz88) 24 days ago
-1 You say " we would need to compromise with consistency if we have availability and partition tolerance. ". This is
↓ incorrect, we don't need to sacrifice anything if there is no network partition, in other words, you don't need to sacrifice consistency/availability unless there is a network partition.
reply

Sign up to post a comment (/users/sign_up/)

Click here to jump start your coding interview preparation (/)

About Us (/pages/about_us/) | FAQ (/pages/faq/) | Contact Us (/pages/contact_us/) | Terms (/pages/terms/) | Privacy Policy (/pages/privacy/)

System Design Questions (/courses/system-design/) | Google Interview Questions (/problems/google-interview-questions) |

Facebook Interview Questions (/problems/facebook-interview-questions) | Amazon Interview Questions (/problems/amazon-interview-questions) |

Microsoft Interview Questions (/problems/microsoft-interview-questions)

Like Us (https://www.facebook.com/interviewbit) Follow Us (https://twitter.com/interview_bit) Email (mailto:hello@interviewbit.com)