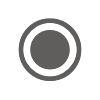
**Transcript**

19 November 2024, 9:26am

 **Anupam Purwar** started transcription

 **Anupam Purwar** 0:03  
So let's let's discuss that. But see one thing is there means right now when you're filtering.  
The information you are only comparing with the relative information, not all the information which is there in the database.

 **Tanmay Singewar** 0:18  
Yes.

 **Anupam Purwar** 0:19  
So it may still so happen that you may have some redundant information.

 **Tanmay Singewar** 0:20  
Yeah.  
Oh, yes, Sir. Exactly. That's why I'd like on the routine check up from the LLM I'm talking about. I will explain it in that solution over there.

 **Anupam Purwar** 0:44  
See, I think you need not worry about that.  
About filtering everything means you can have a separate script which runs from time to time, which will remove redundant information.  
Against each category.  
Because you can have so and that one can leave your real time bandwidth, OK or compute bandwidth.

 **Tanmay Singewar** 1:04  
Yeah.

 **Anupam Purwar** 1:11  
That you can do, say once every day, OK, that you run that script and any information which you find is in there at this database against each category you just drop that.

 **Tanmay Singewar** 1:25  
K when's the 12 best script? It will be a separate complete separate system for the web communicate with the data.

 **Anupam Purwar** 1:30  
It will be a script OK which will just run on your editing.

 **Tanmay Singewar** 1:34  
OK.  
I will just can present.

 **Anupam Purwar** 1:41  
Take it. So this is what I do. OK? From my side. OK. Like how to philtre out the real time? Maybe it is. It is continuing some time. We can.

 **Tanmay Singewar** 1:53  
Uh huh.

 **Anupam Purwar** 1:56  
Because.  
So it is taking our time in the prompt itself.

 **Tanmay Singewar** 2:03  
Oh, honestly though, are we? Was it still high naked joke like trying to restructure entirely in real time or restructuring?

 **Anupam Purwar** 2:04  
Because you.  
So what I'm saying is you can take this out of your prompt, OK, that it says you will save in four years or no. OK, because if you're letting the this is happening in real time, right?

 **Tanmay Singewar** 2:23  
OK.  
The saving information.

 **Anupam Purwar** 2:29  
This categorization is happening while it is generating the answer.  
So your response time would be higher because of that.

 **Tanmay Singewar** 2:38  
Oh, OK.

 **Anupam Purwar** 2:40  
This you can have as a standalone script which will run means you will call you want to use LLM to do that check. It's fine. OK it's OK you enter all the category value pairs OK and then then through LLM OK and then ask the LLM like which are those against the categories. Say say general OK what are what all which all values are actually redundant.  
Or repeat it. Take you can look that you can check their semantics or you can just do a word by word, company whatever comes to your mind.  
And if there is means that could be like a like a what to say like a redundancy module, OK.

 **Tanmay Singewar** 3:32  
Redundancy elimination model, yeah.

 **Anupam Purwar** 3:38  
So redundancy reduction model can be like a standalone code. So are we?  
Yeah, so this is clear, right.

 **Tanmay Singewar** 3:51  
Yeah.

 **Anupam Purwar** 3:52  
No, no. Complete your thoughts. OK, what were you saying?

 **Tanmay Singewar** 3:56  
OK so my screen is visible.  
OK, so first of all, this is the user chat module. So let's say he is talking about his like Father. Like, what's your father do? That's general question we asked and whatever. What for first we will do is get this question and give it to the retrieval category extractor relevant category extractor. So we will go through the LLM like Lama 3.1.  
And give it a such categories which is stored in our DB which has been structured manually structured by us.  
So what he will do is try to find such categories where this question information relies. So he's like we are talking about his father. Like what your father's do. It might have information in the relationship because because it might consists of the family and friends information. So LLM will respond with the relationship tag and that tag is important to retrieve the information from the registered. This is the tagged information.  
So what?  
Next, what we will do is get the tag, get to the Redis.  
And so it's there. Like what? Whatever the keys having the relationship tag get me.  
When the red is what we are looking at it it having the information but it's it's all related to his mother relationship tagged with his mother. He's doing business at Karol Bagh, what he's doing his business of the Handcrop shop is there. His mother is entrepreneur. And what products are there selling selling the embroidery suits and Lena Choris as a traditional Delhi side products. OK, so this is the relative information but it is not relevant to his father.  
OK, but what? This is the relevant information because it's it's from mother's information, which is married to his father.  
And that is a family information. But now what I will do is we have a person of certain like this is a bot where you are from Delhi and I'll the information of the bot is there entire person and other geographical informations. And we have this relevant information in forms of tags. We will give both the information and generate a response from the LLM. Now what will happen is there is a.  
Like specific person of the bot.  
Rate information is there met the father information is upset. So what this LM will do is based on this related information.  
It will extract this information and try to come up with something which is reasonable from the above relevant information as well as his personality or the persona. OK, the information is completely new nuance to the system, but it is relevant.  
OK, so he might come up with a. OK. My father is the manager for my mother. He's come up from the LLM. It has not been stored anywhere but is relevant information. So what happened is we have two tags. The response it will be generated as Jason response. And what happened in the response. It will be given that. Yeah. My father is for the manager for my mother and so on. So information is there and this is the save tag.  
And this save tags will be saved, respond yes and why? Because because this information is completely new based on the previous information.  
If it is already present, if the question is correlated to his mother, like what business do your mother do, then it will respond with the no, because there's a relative. This information is already in the relative information tags. OK, so at this point this response will be sent back to the user at the same time as the. As this second response the second response came from the LLM like that is Lama 3.1 in our case.  
But in back around if this tag is yes, then we will again call one function behind the scenes. This is already sent back to the respond is already sent back to the user and there there is A tag a new function called because the information new and tagged says yes.  
And those response will be to the LLM again and B and we will give back.

 **Anupam Purwar** 8:18  
So yeah, so see because you're able to retrieve it so his mother's information based and that got passed into the LLM as relevant information.

 **Tanmay Singewar** 8:24  
Uh huh.  
Exactly.

 **Anupam Purwar** 8:29  
So it's with LLM made-up something. OK that my my father is a manager to my mother. Right? And then you'll store this in your DB. OK, so that next time when something is asked about the father, OK, it may look in the DB and it may not manufacture somethings again.

 **Tanmay Singewar** 8:35  
Yeah.  
Yeah.  
OK.  
Exactly. Can you learn macroe the same information will be retrieved from the database like this is the mother's information. This is the father's information and same information we carry out to the user. So there will be no like confusion over there. And the memory is sustained.

 **Anupam Purwar** 8:49  
Right.  
That that part is clear to me, OK, what about means you're also saying something about like, how human works? Brain works. OK and.

 **Tanmay Singewar** 9:07  
Yeah.  
OK, that's the part there that says a rearrangement of the same. I just added this block. So what happened is now completely like this flow is completed. The all information is stored in the DB. But what happened is right now it might happen like this, business names will be Co exchange with the mothers and the fathers information. But so it. So it had happen like relationship Mother business is there and he runs a handicap shop.  
And relationship Father business is the manager to his mother. And again there is a run handicap shop and this is the redundant information. But while doing this rearrangement, what we will do is we will come up with some new tags and new arrangements. This DB we have N number of information is there so each tags will be different because this tags are individually generated. Mother's information is generated by some other request.  
Father in solvation is generated by some other request might have a redundancy in there. So what we will do get those tags all tags together and give it to the strong LLM. Like a bigger model might be rgbt 4 and this this information and the what the GPT 4 will do, he will come up with some new tags like we have. And right now we have relationship mother tags.

 **Anupam Purwar** 10:14  
Mm hmm.

 **Tanmay Singewar** 10:35  
Now the Lldm will might we arrange some prompts here and he will come up with a relation. Mother. Father, business. This is the third tag. This is a new tag.  
This is been rearranged. OK, so next time the information's coming rather than giving like 10.  
50-60 keys at a time. He will use less keys like for 5-6 keys.

 **Anupam Purwar** 10:57  
Dude, this is OK. I got it. So you are talking about enriching the information which is already there in the Redis DB, right? Using a LLM.

 **Tanmay Singewar** 11:04  
Oh yeah.  
Exactly.

 **Anupam Purwar** 11:08  
So see, see, I think two things.  
It should happen independent of me.

 **Tanmay Singewar** 11:17  
Mm hmm.

 **Anupam Purwar** 11:19  
What is the redundancy and another is the enrichment, so there should be one enrichment model and one redundancy model.

 **Tanmay Singewar** 11:30  
Mm hmm.

 **Anupam Purwar** 11:31  
OK, so redundancy model will philtre away any repeated information against a category and enrichment module will try to structure the information better and probably come up with some more categories.

 **Tanmay Singewar** 11:47  
Exactly. And those categories we will feedback to the this prompt again means it might happen. Yeah, like it might happen. This categories are not enough.

 **Anupam Purwar** 11:52  
No, this this will also become parameterized.

 **Tanmay Singewar** 11:57  
Yeah, this is also parameterized there.

 **Anupam Purwar** 12:00  
OK, what else?

 **Tanmay Singewar** 12:03  
So this is the entire system, like if it is so then it will surely decrease the size of the DB and at that certain limit it might happen like this new information exchange will not happen anymore. It might happen rarely because new information is will not create again and again.  
So like memory is so saturated right now that all the user queries will be resolved by this already existing formation as we don't create the this memory part the all this user, how the users will communicate on this basis the memory has been built so like it's a human tendency like most human ask for a similar kind of information to the each of them like we use for the forums and other things. They talk about the culture and other things. It's rarely gonna happen. It's new information for the.

 **Anupam Purwar** 12:24  
Mm hmm.

 **Tanmay Singewar** 12:55  
This might be a saturation level, and at that time our redundant system or disclosure system is commonly so strong and we can rearrange some of the tags and we can go for the manual work also.  
Speed up the entire process.  
Each board.

 **Anupam Purwar** 13:13  
Yeah, I did not get this laugh part.

 **Tanmay Singewar** 13:16  
OK, this is the like. This is saturation level means new information will never be created. Rarely been created so OK so there's A1 user we joined recently and we have no information. He continuously talks about something, something, something something and we have generated the memory like small amount of then 10 user joins is us and they come up with some new new questions and new queries and this will happen in us like large amount like let's we have 100 user now.

 **Anupam Purwar** 13:24  
Mm hmm.

 **Tanmay Singewar** 13:46  
But like after the 100th user, the user are asking similar type of questions again and again. Those information is already in the Redis. The relevant information already stored. So for those users we don't have to create a new individual type of memories because it is already there. So it it will grace to the certain bar where new information rarely creates.  
And at that time, we will restructure the entire DB manually or with a higher class means the we have generated information so classified that it behaves like a certain person.  
And so it was so good at enrich database.  
This entire.

 **Anupam Purwar** 14:32  
So now what you're saying is?  
Over a period of time.  
Means you won't have any more information to be really filled in in this DB because.  
Through all the interactions and hundreds of people.  
Like it would have already stored the information there, so it it means even if you if there first of all it will have the information against the category so.

 **Tanmay Singewar** 14:54  
Yes.  
Mm hmm.

 **Anupam Purwar** 15:06  
And 2nd like when you will suppose you even end up storing something again, OK.

 **Tanmay Singewar** 15:13  
OK.

 **Anupam Purwar** 15:14  
Then your redundancy mechanism will catch it and not let it get stored, because search information already exist.

 **Tanmay Singewar** 15:20  
Yes.  
Exactly.

 **Anupam Purwar** 15:25  
In that way your address DB will stop growing in size.

 **Tanmay Singewar** 15:28  
Yeah, that will be a like fixed size. We can have a certain amount, yeah.

 **Anupam Purwar** 15:31  
I think it this is fine.

 **Tanmay Singewar** 15:35  
Good answer, this is the issue like you are tackled. I want like we have coming up with these two tags. So I like your like approach. We will just give the response and 3rd call will be always be generated at any cost it's for the right now because yeah this is a tagging system I come up with. So third call will may or may not but we will generate the third call because for initial stages it is good for us.  
So in this we will not say in in say in four or info as soon as this response comes, we will send back the response.  
And we will so extract try to extract it from the. We will change this prompt or where it is like if it is in the formation, do not extract the text or the leave the here or it is there. If it is information is not there then extract it and send to the Redis this with this entire saving information task we will try with another.

 **Anupam Purwar** 16:33  
The other way you can do it like you should return the response OK and means let that interaction with LLM continue.

 **Tanmay Singewar** 16:42  
Honest about that, that that might happen. Like if he he gets certain time after sometime he might come up with a again this similar type of question and LLM will respond with a new information again so that might create something that is.

 **Anupam Purwar** 16:58  
Anyway, what I'm saying is.

 **Tanmay Singewar** 17:00  
OK.

 **Anupam Purwar** 17:01  
Wherever you go, get again written the save info tag. OK right now you're generating it while generating the answer itself.

 **Tanmay Singewar** 17:10  
Yeah.

 **Anupam Purwar** 17:11  
TK TK means.  
What I'm saying is don't do that, OK, in this prompt you make another call to LLM, OK?

 **Tanmay Singewar** 17:23  
OK.

 **Anupam Purwar** 17:23  
3rd call to the LLM but not within this prompt. OK.

 **Tanmay Singewar** 17:25  
Mm hmm.  
Oh, OK.

 **Anupam Purwar** 17:29  
Once it it has generated the answer.

 **Tanmay Singewar** 17:32  
Uh huh.

 **Anupam Purwar** 17:33  
OK, you make another call another function call and then put this prompt. OK well to check whether this information is there.  
In the relative information or not?

 **Tanmay Singewar** 17:49  
OK, now he calls.

 **Anupam Purwar** 17:51  
Ticket means right now you're doing it all within the same prompt.

 **Tanmay Singewar** 17:54  
Yeah, exactly.

 **Anupam Purwar** 17:56  
So it is increasing the latency and your user should get the response immediately. Here it means Avi. What is happening is.

 **Tanmay Singewar** 17:58  
Oh.

 **Anupam Purwar** 18:06  
It is. Can you send me that? Have you uploaded that code somewhere?

 **Tanmay Singewar** 18:13  
The code is.

 **Anupam Purwar** 18:18  
If you upload the code now, I will edit it and show you.

 **Tanmay Singewar** 18:21  
OK.  
It's something like CV chat.

 **Anupam Purwar** 18:28  
See you in that.

 **Tanmay Singewar** 18:29  
Yeah, SCVMM. This is the model. OK, I'll just for now. I will just make it public or.

 **Anupam Purwar** 18:37  
No, no. Just push it to the strap. Oh, no.

 **Tanmay Singewar** 18:41  
OK and.

 **Anupam Purwar** 18:41  
No need to make it public.

 **Tanmay Singewar** 18:47  
Is it?

 **Anupam Purwar** 18:50  
Upload download.

 **Tanmay Singewar** 19:59  
Other was it OK?  
Yes, I'm done.

 **Anupam Purwar** 20:37  
In which folder have you applied?  
So you have product in a stand alone.

 **Tanmay Singewar** 20:40  
Oh yeah.

 **Anupam Purwar** 20:43  
Take care.  
No, I'm not able to open it.

 **Tanmay Singewar** 20:54  
10 metres are on this.  
Give it a model one. I think it's responding to the repo.  
Yeah.  
The model and main file is there in the repo now.

 **Anupam Purwar** 22:40  
OK.  
So this is actually a notebook file so.  
And cannot edit it here properly so see what I am saying is.  
See. You hear Joanna. Till tomorrow. Whatever response that you are generating, that is fine.

 **Tanmay Singewar** 23:30  
OK.

 **Anupam Purwar** 23:30  
The same for Vala joke you are giving.

 **Tanmay Singewar** 23:34  
Uh huh.

 **Anupam Purwar** 23:35  
You get that you can do in the next call.

 **Tanmay Singewar** 23:41  
OK.

 **Anupam Purwar** 23:41  
I did return the response.  
And you take that response and show it on the render it on the UI.

 **Tanmay Singewar** 23:48  
OK, so now there is a two compulsory call has been generated. 3rd is a conditional call and we will now we will switch back to three, three compulsory calls.  
The llms as then conditionally we will generate this let me.

 **Anupam Purwar** 24:04  
That is fine. Means let let there be a simple recall. I think that is fine, but this will save time. It will reduce our response time.

 **Tanmay Singewar** 24:08  
OK.  
Oh yeah, sure, sure. So.  
We'll produce this other than conditionally, it will compulsory make the third call check the check the information, whatever it is. If it is a new, throw it into the DB or like creating those stacks else or drop the call.

 **Anupam Purwar** 24:20  
Hmm. Take it.

 **Tanmay Singewar** 24:33  
Yeah.

 **Anupam Purwar** 24:34  
Because I just updated that accordingly. What else you were saying that that about getting motivated from the graphs?

 **Tanmay Singewar** 24:45  
For those purpose, like what the graph actually doing is, they are tagging it information like previously. What is the like Microsoft RAG is doing graph rack is doing, they're taking the information converting into such a way that it can be representable in the graph, but what we are right now we are doing is that if they are converting into the rack graph information, we are converting into A tag key value pairs in the Redis. Redis is faster and it is simple to manage because it's just a key value.  
Yes, you don't have to manage another skill set to manage those graphs, because graph is another cluster management is there.

 **Anupam Purwar** 25:25  
Mm hmm.

 **Tanmay Singewar** 25:26  
And we can switch it back to the things but, but this will be totally inspired by this graph, right? Rather than storing each and every information in the graph branches, creating the branches and other things, we will creating a simple where the smaller models like we are using this Lama 3.1 and this graph required like higher GPT 4 level intelligence right now. But we are using smaller and concise model. We are prompting in well.

 **Anupam Purwar** 25:49  
Mm hmm mm hmm.

 **Tanmay Singewar** 25:56  
That way, and we can come up with these those tags and in the graph system that restoring it is another new level task because we can't restructure anytime sooner because if the trees are been created.

 **Anupam Purwar** 25:59  
Hmm.  
Because it it is, it has a hierarchy. OK, so you.

 **Tanmay Singewar** 26:14  
Yeah, had a case.

 **Anupam Purwar** 26:16  
There's multiple layers of hierarchy you have to fit it in that hierarchy.

 **Tanmay Singewar** 26:20  
Exactly. But if the information is returned, information is there and there is a like another branch is there. We then can't cut out those branches. But here we can.

 **Anupam Purwar** 26:31  
Right. Well, because neither you have to cut the.  
Start cutting from the upper node.

 **Tanmay Singewar** 26:38  
Oh yeah, if there is redundant and if this is like if this branch is having a 60% new information and 40% redundant and other branches having like 80% new information and 20 this like 40 and 20% will be information there only because we can't cut out those branches because the new information will be cut down.

 **Anupam Purwar** 27:00  
Hmm. Hmm. Hmm, because it would have been nested under the under the.

 **Tanmay Singewar** 27:02  
Correlation with those, yeah.  
Exactly though.

 **Anupam Purwar** 27:08  
You know, I get that.  
So we need to have a redundancy module and.  
In return module.

 **Tanmay Singewar** 27:16  
Embracement. Yeah, to models divided.

 **Anupam Purwar** 27:18  
We can suggest. Note this down, OK? Add it in your diagram itself. OK, that we have a redundancy and this module and this will run OK independently.

 **Tanmay Singewar** 27:23  
So.  
Exactly, yeah.

 **Anupam Purwar** 27:33  
OK, you need not run it along with the.  
In real time when you are generating the response.  
When I get means whenever the traffic is low, say the, you can trigger that script whenever the CPU usage OK of your system is low.  
So you can write a. You can write some logic OK using OS commands whenever the CPU usage is low, then you trigger that script.  
Right. So in that way, do you think there are times when your CTP usage is low? OK, but not many people are coming and chatting that time. We are doing the dependency check and.

 **Tanmay Singewar** 28:15  
Down.

 **Anupam Purwar** 28:25  
Enrichment.

 **Tanmay Singewar** 28:27  
Yeah, this will be manageable by the database because database also required some amount of pooling, pooling, request pooling request or the bandwidth required so that we will just fasten the process.

 **Anupam Purwar** 28:37  
Right.

 **Tanmay Singewar** 28:41  
Database also have whenever it is cool down period for the database we will pick up some sort of information and addressment model and redundancy elimination model will be start working on it.  
Yeah.  
Yes, Sir. This I will start working on this as well as this UI part also, but because the development part and this is in rearrange because changing the modules there's 2-3 parts I will update it tomorrow only today itself and the changing the modules or change by changing the like llms also because 3.1 is good in or. I tried with Lama 3 but it is not that good.  
At this part because following the instructions that what I understood 3.1 is good, I I haven't tried 3.2 they just released the 3.2 also.  
Now.

 **Anupam Purwar** 29:36  
Hmm, maybe we will stick with the Islam models only, OK?

 **Tanmay Singewar** 29:39  
Yeah, exactly.

 **Anupam Purwar** 29:41  
And then we like.  
When you just 70 or 80 billion type of models.

 **Tanmay Singewar** 29:47  
80 billion is good because this this will also because this categorization required some of amount of previous informations because we are giving some less informations here like skills and general. But in general we have a general information we want to what general specifies that should be come up from the LLM.

 **Anupam Purwar** 30:31  
Hmm.

 **Tanmay Singewar** 30:31  
Lamo, 3.1 ADP because it just waving this relevant information and trying to respond on it based on its persona, so that this way also we can.  
Decrease the like.  
Response time because LL smaller LM faster response.  
Two different task and two different llms will be.

 **Anupam Purwar** 30:56  
No, it'll cool. I think. Yeah. That's how you should.

 **Tanmay Singewar** 30:58  
Yeah.  
Exactly. And while doing this, because this is a once in awhile task we can use like bigger models here like.

 **Anupam Purwar** 31:08  
But there there we can use bigger models, but I could we add that OK use use. So we're gonna LLM.

 **Tanmay Singewar** 31:10  
Yeah.  
OK here.  
Yeah, Sir. That's it. Oh, you're just. You're sleep. We'll try our also use, like, stick to the Lamaya family who can use, like, oh.

 **Anupam Purwar** 31:44  
Tell me also about this Redis.  
DB work where where are you offering this and how? How did you create it?

 **Tanmay Singewar** 31:52  
OK. Does this red part is actually.  
I.  
Where I hosted OK.

 **Anupam Purwar** 32:04  
It's unstructured database.

 **Tanmay Singewar** 32:04  
I think.  
This is a structured turn off. This is like caching system. This is not a like so persistent storage system. This is a caching system where we have to use certain.  
External DB also here.  
But this is a quite fast because it stores its entire database in the RAM or the it's primary storage.  
That select this is a good part here.

 **Anupam Purwar** 32:33  
Is the external DB that you are using.

 **Tanmay Singewar** 32:36  
Oh, no. Right now we are not using external as we like development site of project. So as we go to the production we will soarly use this external DB to store those information.

 **Anupam Purwar** 32:50  
So this redundancy and enrichment module, if they have to run, you have need to have this information in some DB right?

 **Tanmay Singewar** 32:54  
Mm hmm.  
Yeah.

 **Anupam Purwar** 33:01  
But like.

 **Tanmay Singewar** 33:01  
That will be some sort of or like consistent like it will be like kind of Postgres or anything.

 **Anupam Purwar** 33:13  
But you're saying here that this loads all the information in your RAM. OK.  
So.  
If you're, let's just show me like.  
Where does that load the information from? Are you to load in the RAM the information needs to be there somewhere.

 **Tanmay Singewar** 33:33  
Yeah, like for the simplicity we have used it in a certain way that all information is created and loaded in the RAM itself for now. But what happened is as it is stored, it's sort of caching system. We will store entire information as a cache and once on a while we will try to like within the two hours, 3 hours or 1/2 hour cycle. As soon as the information information is created.  
We'll take out this information and store it in a so like personal storage, like some sort of. We might store it into the.

 **Anupam Purwar** 34:06  
Are we are we tell you one thing, it's.

 **Tanmay Singewar** 34:08  
Uh huh.

 **Anupam Purwar** 34:09  
Is it still happening that when I'm chatting with it?

 **Tanmay Singewar** 34:14  
OK.

 **Anupam Purwar** 34:15  
The information is there in the RAM OK, but after 30 minutes if I start chatting again, that previous information is lost.

 **Tanmay Singewar** 34:17  
Yes.  
No, no, no, not right now because we are using the Redis as a third party service. That third party service actually stores entirely because Redis is storage, but it stores in the RAM means if the computer restarted only that time, the information is lost.  
It's not social like that. It's not gonna. It gonna use some sort of storage or like a heap system where it stores the information strongly retrieval. Trying is like we have to use some low time. That's why you are using the Redis because there is in the.

 **Anupam Purwar** 34:59  
So you have to right now suppose I'm chatting right now TK and it is storing that information across categories, right?

 **Tanmay Singewar** 35:02  
Mm hmm. OK.  
Mm hmm.

 **Anupam Purwar** 35:11  
Bono log means suppose I asked about his occupation and education.  
This information was not there. It made-up something like and then after making it up, it is stored in the Redis ticket Redis DB. Tomorrow again I ask the same question. Now what will happen?

 **Tanmay Singewar** 35:26  
Yeah.  
Yeah, the cache will be same as the previous one because it's a managed service, so cache will be stored as. There will no persistence to like no caching like flush out.

 **Anupam Purwar** 35:47  
But tomorrow's system?

 **Tanmay Singewar** 35:48  
There is a TTS. Oh yeah, there is a TTS system time to leave or try to time to leave TTL system means we can set it. There is a default ttls also, but what I'm it is after this TTL this cache is flushed out. But before that Flushing we will store this entire entire data because data in some sort of secondary storage.

 **Anupam Purwar** 36:12  
So what is that secondary storage right now for you?

 **Tanmay Singewar** 36:15  
No, right now there isn't no second storage. We will come up with some sort of any database.  
We will use certainly the Postgres means there will be another section here.

 **Anupam Purwar** 36:26  
And what is this Redis cloud?

 **Tanmay Singewar** 36:30  
Yeah, Redis Cloud is a like it's a service. Redis is a in memory storage DB and Redis cloud is a service. They can provide it like enterprise system.

 **Anupam Purwar** 36:42  
So, Amir, you're saying there's time for it to get flushed? So is it getting flushed out currently?

 **Tanmay Singewar** 36:47  
No, not not getting sooner, not getting sooner.

 **Anupam Purwar** 36:53  
Tomorrow time here. What is the time till what time will it start?

 **Tanmay Singewar** 36:57  
Yeah. OK. For now we are using this free service that free service let us.  
That, like you have a certain like 1GB2GB data is storage system is there after some amount of time it will just full it will start to eliminate it.

 **Anupam Purwar** 37:18  
And who is providing this 1GB space?

 **Tanmay Singewar** 37:21  
Ki am using this service name. I have to look for it. I just got this is free service.  
OK.

 **Anupam Purwar** 37:32  
Some persistent storage right right now.

 **Tanmay Singewar** 37:36  
Right now, no, no. We are using Redis only Redis, no processor storage is there. It might happen that data has been flushed.  
But it will not, because this is a service. OK, I don't.  
In flash, OK.  
OK, this is the one. This is the server's data platform where they provide. This is the Redis service also.  
This is the storage. OK, there is a request like 1000 requests per day. They're providing bandwidth of 50 GB storage of 256 MB only.  
As this MBS person, like this amount of information is out of the their subset, then they will start flushing out the information.

 **Anupam Purwar** 38:35  
Me that I get OK so.

 **Tanmay Singewar** 38:36  
OK.

 **Anupam Purwar** 38:38  
I'm not clear. OK. OK. How are you? So this information is going in this upstage to 256 MB.  
We chat.  
Every corresponding information re. It is pushing it into this 256 MB.

 **Tanmay Singewar** 38:54  
Mm hmm.  
Exactly.

 **Anupam Purwar** 38:59  
Provided by.

 **Tanmay Singewar** 39:02  
Yeah.

 **Anupam Purwar** 39:03  
And that is you have. It is like a layer over this information which loads this into the RAM.

 **Tanmay Singewar** 39:12  
Redis entirely of system is. This is the they are having their system and their ram we are using.

 **Anupam Purwar** 39:20  
Sorry, come again.

 **Tanmay Singewar** 39:20  
This is the host statements Redis. This is the service provider up stash. They're providing the service as Redis as a service, so means they are hosting some some sort of machine and in that the Redis is running and Redis what is doing is it will have in their ram in their system we will having our data but the fetching is so fast.

 **Anupam Purwar** 39:26  
None.

 **Tanmay Singewar** 39:45  
We are only having the latency of the bandwidth.

 **Anupam Purwar** 39:49  
Yeah, ram. Tamari is it of this upstash or is it the ram of your EC2?

 **Tanmay Singewar** 39:55  
No, no, it's up top up upstairs.

 **Anupam Purwar** 39:57  
How to then then can you make this also clear in your diagram?  
Well, like one may think, OK, that this is the RAM that you're consuming all the same instance where your LLM code is running.

 **Tanmay Singewar** 40:13  
If we might, we can use it before for running the Redis. It might require one or two number of CPUs. So as we have like for the simplicity I use the third party service but we can use this entire Redis system in our in our like.  
CPU instance, but that will really help because of it will be in a nanoseconds retrieval. Right now it might take a one or two, or it may it's taking right now like a 40 or 50 milliseconds.  
But if we come up with a larger, larger number of like T2, mini, medium or something where we can host our entire system on our own, just download it, it will surely it will be in a nanoseconds. Because if when our trying from my local system I created this entire like downloaded the Redis in my system.  
And run this all in text. The minimum response time was the 900 milliseconds 1 seconds 1.21.3 that was highest I am getting.

 **Anupam Purwar** 41:21  
Mm hmm.

 **Tanmay Singewar** 41:22  
Whatever the query I was you are throwing.  
For for the like, but we are running in the development server. We are using this managed service from the OPS dash we are getting response like 2 seconds, 3 seconds 4 seconds. Sometimes I think that's a latency issue from this section because there is a bad bit latency.

 **Anupam Purwar** 41:41  
Mm hmm.

 **Tanmay Singewar** 41:42  
Yeah.

 **Anupam Purwar** 41:44  
OK.

 **Tanmay Singewar** 41:45  
Yeah, exactly. And they're having a large throughput. That's the added advantage means while having this persistent storages where like a Postgres and other things they might having certain number of limits because at at that time at this time there will be no input is there or no output is there there will be there. But in the Redis what happened is they're having a large throughput and there is a number of amount can be inserted and amount can be.  
Taken out.  
So they handles a pretty large number of data and that's why the chat system also required this. So we are using the disc.

 **Anupam Purwar** 42:25  
Hmm, So what you are saying is this this Redis service? You can also run in your EC2 if it is large enough.

 **Tanmay Singewar** 42:31  
Exactly. No, not that large. You know, it required a more like 2-2 hundred, 300 MBS of RAM and one CPU.

 **Anupam Purwar** 42:41  
Tiger. But right now, in place of and and in that case, the RAM that it will leave to load the data would be of the TC2 where this Redis service is running right now. You are using some other platform of stash.

 **Tanmay Singewar** 42:51  
Yeah.  
Yeah.

 **Anupam Purwar** 42:58  
Where this Redis service is running and the information is getting loaded in an in and from your fetching the information.  
When whenever you are waking up, call.

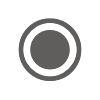
 **Tanmay Singewar** 43:11  
Yeah, exactly.

 **Anupam Purwar** 43:13  
Other than then outline this alternative low diagonal.

 **Tanmay Singewar** 43:17  
Yes, I was so.

 **Anupam Purwar** 43:20  
It was to me it all like a this is also installed within our EC2 only this. This was my impression but PK you you add those details.

 **Tanmay Singewar** 43:26  
No.

 **Anupam Purwar** stopped transcription