**Annex cloud-Kanini Brainstorming Workshop 2-20240326\_082827-Meeting Recording**

March 26, 2024, 3:01AM

2h 15m 12s

 **Speaker 1** 0:26  
So for today's session, I think let's first get out of the.  
Wait, what?  
We have to do it later in the day, right?  
So, Prince, your did paying about the portability, right?  
I think I agree that looks like a good option so we can start working on seeing if we can take the data from using Lambda function that that is something that we'll do later today.  
I think the aim is to get the ohh the busier man.  
And Gaffa set up so that we can at least and we to him, let me know if I'm right.  
Like we are going to 1st put the data into S3 bucket right think will be S3 bucket that is what we have to get out of there right correct.

 **Tuhin** 1:18  
So I think let's do this.  
Let me go ahead and share my screen.

 **Speaker 1** 1:24  
Umm.

 **Tuhin** 1:24  
Let's review the plan and see where we are.  
I think gonna high level that makes sense.  
Priyanka, right.

 **Speaker 1** 1:29  
Yeah, yeah, yeah.

 **Tuhin** 1:31  
Uh, But let's go ahead and.  
Translate.  
You know one second.  
Alright, ohm.  
So centrally Priyanka, I think uh your team should have us have access to a defendant to add a new one.  
Please let me know.  
I'll go ahead and add them. Umm.  
Also to the worksheet that we started, let me.  
Let me pull up that worksheet.  
That had all of the steps.  
In the answer them this, sorry.  
And you see my window still OK.

 **Speaker 1** 3:45  
Yes.

 **Tuhin** 3:46  
OK, so I think this is where we started.  
And just to uh, we started transferring the plan over to this Monday board so we can.  
You can, uh, work together a little bit.  
If we come in here, so obviously we started talking about.

 **Speaker 1** 4:12  
Umm.

 **Tuhin** 4:13  
Uh.  
Talking about setting up.  
Paul, let's see the data model.  
I think Sachin has already started.  
We can talk through that.  
Verification and transformation and the Lambda function.  
So this is the second portion, I think Priyanka, you're talking about it here.

 **Speaker 1** 4:39  
Yes, that is a that is something.

 **Tuhin** 4:39  
This. This I.  
Yeah, on this item here. Umm.

 **Speaker 1** 4:45  
Umm.

 **Tuhin** 4:47  
So maybe we run through it Karthick this item of creating my sequel, RDS My sequel A.

 **Speaker 1** 4:52  
Umm.

 **Tuhin** 4:58  
This is done right already.

 **Karthik Raghu** 5:02  
Yesterday that is done.

 **Tuhin** 5:09  
So created technical worksheet so.  
I believe this is.  
In progress.  
We have not.  
Umm pleted that?

 **Jubaed Prince** 5:29  
Well, we do have our draft for it and I was under the workshops will go through this workshop.

 **Speaker 1** 5:40  
I also thought we are going to go through this, right?

 **Tuhin** 5:41  
Is this?  
That is, is this the draft or or who?

 **Jubaed Prince** 5:45  
Yes.

 **Tuhin** 5:46  
One is the watching, OK.

 **Jubaed Prince** 5:48  
This is around, yes, the A2M.

 **Tuhin** 5:55  
8 to one.  
OK, so I know Sachin has, so maybe we just run through this.  
So Sachin has already created a Sachin, let me try to.

 **Jubaed Prince** 6:02  
Umm.

 **Tuhin** 6:09  
Find the.  
The worksheet daddy you have started and I didn't think.  
Any has access to this, so I'll go ahead and I'll start share.

 **Sachin Naikwadi** 6:18  
Yeah.

 **Tuhin** 6:24  
I'll start sharing the access and uh, do you wanna take off?

 **Speaker 1** 6:27  
Umm.

 **Tuhin** 6:30  
Uh, Sachin, and walked through this and I'll just stop sharing and I'll, I'll go ahead and I'll, I'll, I'll give the access to kanini folks on the background.

 **Sachin Naikwadi** 6:33  
Yes.  
No.

 **Tuhin** 6:40  
Then we'll come back.

 **Sachin Naikwadi** 6:41  
OK.  
Yeah.  
Let me see your message.

 **Speaker 1** 6:47  
So in the mean uh session sharing, I think, uh.  
I was talking about uh, this establishing the the vision connection and data sync mechanism, right, Karthik, we can reach out to you later after the meeting and we can work on that, right?  
Even your team.  
There's gonna be as possible they.

 **Karthik Raghu** 7:07  
Good.  
We already have the setup.  
Uh, we can.

 **Speaker 1** 7:12  
Awesome.

 **Karthik Raghu** 7:12  
Yeah, you can work with me on that.

 **Speaker 1** 7:16  
Sure.  
Thank you.

 **Karthik Raghu** 7:17  
Pink to him like I think when we were discussing, we told that this is not the final one.  
I think the idea was more on the proposal standpoint wherein what you guys think would be appropriate, right?  
So I think may need to propose it, which I was just asking is like whether we want to push things into S3 bucket.  
I think that design is as far as I know, to.

 **Speaker 1** 7:44  
Sure.  
He so you have to.

 **Tuhin** 7:53  
He yeah.  
I think there was a there was a question mark on the three, three back that I think there was a A, there was an open discussion if I recall right.  
I think there was a divergent in opinion, right?

 **Karthik Raghu** 8:05  
This.

 **Speaker 1** 8:05  
Is.

 **Tuhin** 8:07  
Yeah, so.  
Just, uh, let's Sachin go through this doc.  
So for the data model.

 **Speaker 1** 8:14  
Umm.

 **Tuhin** 8:15  
Then once we come back to the.  
Uh, the debezium and the Kafka stuff and the and the Lambda function or whatever function we're gonna application we're gonna create for CDC.

 **Speaker 1** 8:25  
Umm.

 **Tuhin** 8:27  
Let's take a deep dive on that, like in.  
Yeah, but yes, Karthick, you're right.

 **Speaker 1** 8:33  
Sure.

 **Jubaed Prince** 8:33  
Yeah.

 **Tuhin** 8:35  
You're Karthick.  
We will.  
Do we did put a pin on it?  
So we have to come back to that.  
Yeah.

 **Jubaed Prince** 8:40  
Umm, just for.  
Add to that since we have two hours, can we spend the first hour in this data model worksheet exercise and then the last hour can be CDC discussion and what path we take?  
How does that sound?

 **Speaker 1** 9:00  
Sounds good to me.

 **Jubaed Prince** 9:07  
Does anyone think otherwise?  
This.  
Ohh umm.  
So, umm, Sachin, are you gonna?  
Walk us through that document.

 **Sachin Naikwadi** 9:22  
Yeah.  
Yes.  
So this is the purpose table structure, so I will go one by one.  
So first table uh yes, 15 V 3 users summary.  
So this will contain a user attributes.  
So you ID is the primary identifier identifier for loyalty, so this may be email address.  
This may be a customer ID anything client wants, so this is the loyalty user identifier.  
Everything identify UID, the name, email address, birthdate and your three date.  
Let me go to this you birthday and you'll see opt in status opting out date.  
Those are user specific attributes.  
We are going to add this thing or summary.  
Uh.  
Along with user attributes, we are going to add points related.  
Column.  
So total credit points.  
How many points he earn?  
So we will add a some in this column.  
Similarly, total debit points.  
All deducted points will sum up here until the redeem points, so it's specifically to redeem.  
That is, uh, our action ID 107 for redemption.  
Or if any user redeem those points on order, they very in points and take on coupon and they use punches.  
So this is a specifically to redeem points.  
Naked expired points.  
Uh, how many users?  
Points expire total, so we'll add here.  
Uh, redemption credit.  
So if anyone redeem points but user cancel that order or return that order.  
So we will giving back those points.  
So those total points, manual debits if any adjustment happen on that user for debiting the points.  
So that's some will go up here, but then manual debits, uh, total return, David's if user return the orders so we have.  
Divided the points which he earned on that order.  
So not only turn David points, we some.  
And total balance funds means uh, he's available points the balance funds.  
So these are the all uh, summer for related points.  
Then actions performed in his lifetime.  
Uh, how many actions you perform like create an account, purchase any custom actions, so all actions will go up here.  
So we will add a comma separated all actions here in left and how many actions he performed.  
Uh, last activated it.  
We will maintain here his last activity date.  
Ohh by the purchase.  
Really meeting. Uh.  
Last purchase date we will maintain here when he did last purchase.  
So we will maintain, yes, two show on .7 mode.  
They knew the credit when he created a user update date and this year syncing credit and updated in table get synced.  
This table or from the Afghan who we can maintain creating update person.  
So these are the columns.  
So how data will flow in this table?  
So let me go to that.  
When any users created we will from Kafka we will look for that bin log and that topic and if you insert happen on user table so we will add that you able also.  
Update will happen when user perform any update on his attributes like name, email address.  
So we have to take bin log and on the basis of that we can update some meeting as well as if you go do purchases, users do rename Zion.  
So if that hybrid happen, we will track that activity from the our syncing script from backup and we will maintain the points if credit points happen, we will add here those points, total checkpoint plus whatever he points out similar for debits.  
If David happen, we will reduce problem.  
This total labels how many points he really uh, similarly, probably expire feelings.  
And here we can maintain those balances on the basis of back to it.  
So specifically as if actions performed, we will add actions here and we will maintain the last activity date last birthday date.  
So basically for this summary, we have to watch these tables user, user detail and transaction dinner.  
So any activity going to happen in this table, so we have to check and update the on this table respectively.  
So this.  
And from this, uh, let me go to the report first.  
Uh second MPB summer.  
So we have a multi point bucket.  
So on one time the action user can earn.  
The regular points as well as multipoint bucket points.  
So for multi bucket points we have created another table and we have maintained the summary so I will go through those fields.  
So basically here user ID we have that bucket IDE for each multi bucket we have a bucket ID as a primary key.  
So on the basis of bucket ID, we will maintain its that bucket credit points and bucket debit points as well as balance points related to two that bucket.  
Uh, for that we have to watch.  
Yes, 15 with three multipoint baggage transaction.  
So any activity happen in this table so we have to ohh think relatively in this MV summary.  
So in our script we have to watch this table for the MB summary application.  
So before going to transaction and I will you first, how report will work if user gets searched.  
So if report is searched by, usually nothing is added for start date ended, just added UID and searched.  
So we have these columns on reports, so name will come from user summary, user ID, UID will come from summary email address.  
Actions performed will come from summary.  
Lifetime points will come from summary total credit points, so we can directly show.  
Uh.  
Similarly for multi bucket points.  
So we have multiple buckets, so don't for me MPV summary we can pull total credit funds by pocket twice.  
So there are four buckets, then the four total credits will display relatively with bucket name.  
Similarly with her own points.  
So all these columns so we can directly page from others user summary or multiple bucket summary.  
So if you are research we can get from these two tables the values.  
So next is if user do perform at a search by date, so we can't pull directly from this, so we have to maintain 2 tables to keep a transaction history with the dates.  
So that's why next table come into picture.  
Yes, 15 we 3 transactions summary.  
So this table let me share the values bills.  
So this is the transaction ID from our ESP with three transaction table.  
So com that we will pull here then inside ID, temperature, ID, user ID, action ID, what action he did perform.  
Uh.  
Points credit or debit?  
How many points he credited or debited if that transaction is related to purchase then order ID or the amount or the discount if it's eight to to remove then reward ID yeah reward code if that order contains reward code Leeward status if that is used or it is redeemed.  
Uh reason a resource to ID experiment for the those points when those points feelings right.  
But transaction credit and transaction updated.  
So in the transaction is updated so, but this table so any transaction happen in yes 15 meter transaction and yes we do need the transaction detail then we have to uh think those changes by checking news detail either insert, update, delete and we have to creatively do changes in this summary table transaction summary table.  
Next it's a multi point bucket transaction.  
For one transaction, there are multiple multiple bucket transactions.  
So similar transaction ID will come here and then bucket ID is the main concern here.  
So bucket ID for which bucket here going to earn create or debit points it or debit points and action I.  
And then for similar expired date for those points.  
So this is a multi bucket transaction summary.  
So for that, we'll have to watch.  
Yes, 15 with the MULTIPLICATE transition, any insert, update, deletion will happen.  
Then we will sync to this MPV transaction summary.  
Uh.  
When 30s happen on start date and end date if search by date range so name we can pull.  
As a action performs, we have to take from that date range.  
We have to use just 15 with the transaction summary where to pull action ID S on the that day.  
Drinks by combat, the distinct action ID.  
Uh, lifetime points.  
We can take directly from somewhere there because we don't have a data ring on there.  
Similarly, multiplying bucket lifetime points one points we have a date range, so from transaction summary we have to pull credit some of credit by date range Emily multiplied bucket on points by bucket ID plus some by bucket ID.  
We have to pull the sum of credits for that data.  
Redeem points.  
We have data in so we have to take from transaction summary.  
Similarly, return points, David points expire points balance.  
We can directly take from user summary as we don't have a.  
Date criteria to search on that.  
And I've been status we can like the pool for summer.  
So.  
These are the tables, so initial migration we will going to create one PHP script and we have we will check all these current tables.  
And we will migrate their data by PHP script to the new data model.  
Uh, simplices.  
We have to keep in CDC, in come Kafka, we're going to read topics.  
The below are the important keypads when created user update user.  
We have to check that event on redeem points in order create cheap return cancer.  
Anything happened?  
Merge user happen GDPR.  
Did it happen?  
It program is explicit without retro.  
Then I'll talk to the user is happening.  
So in October user we have to delete the users history so we have to.  
Keep in mind that are In Sync.  
Posted this also gets covered.  
Uh, about the partition.  
One thought is that uh, the transaction tables so.  
MTB transaction and transactions.  
So this will contain a lot more data in future.  
It's going to a bigger table, so I'm thinking we should have a partition.  
So this is the just thought.  
What do we have to check?  
So we have applied 1 partition and month plus year.  
So for each month there should be 1 partition.  
So fetching data should be the faster, so that is the one thought process.  
So in future this will not the slow down the process, so we have to change that option.  
Anything question here?  
Yeah, I'm a little bit.

 **Speaker 1** 24:31  
Yes, I think I wonderful.

 **Tuhin** 24:32  
That's it, yeah.  
So thank you Sachin, for walking to this.  
I know this is some amount of information to probably digest and go through and review.  
But Priyanka in terms of.  
So the way I said, maybe it's simplistic and searching you can keep me honest here too, right?  
So Sachin's kind of proposing.  
This table structure to be in place for the points report and if we have the base tables or the triggers, what?  
Is identified right?  
Who which is going to make the updates to this table and like whether that's an insert, delete or updates?  
And Sachin, I think if we likely have the queries already, again at least come up with the sequel queries that is going to be necessary for this updates too.

 **Sachin Naikwadi** 25:28  
Yeah.

 **Tuhin** 25:45  
Or or at least the logic for the update stalker in this destination tables right?  
Is that fair to say, Sachin right?  
Even if we don't have like say the queries or what have you, but we we we do know that logic right based on the situation that how how how this has to be transferred over.  
Then Priyanka, I think is that enough for us to.  
Take the next next step in terms of thinking about like you know, how do we take the Kavya events and essentially send it to the destination table?  
Or part would be involved from your point of view.  
Like you know, for other some of the additional questions that he may have get to that state.

 **Speaker 1** 26:39  
So in of one thing I was thinking was going to what's happening is when doing right, uh, we have this what indexes needed to go to like update what?  
But they are all going into the same table, right?  
And how frequent is that change?  
And we actually want to read entirely Anand.  
Ashok, please feel free to jump and head is my.  
My thinking is that if we go by column wise and the data is being updated in one column at one second and another column in another one, is that how we can set it up because there will be a lot of parallel processing or the entire record if it's updated we calculate everything based on that based on new ID and then Prashant you guys have any.  
Alternate.

 **Anand sivaraman** 27:32  
Can be the approach can be 2 fold Priyanka so when you know when you say columnwise updated.

 **Speaker 1** 27:32  
And yeah, you heard that.  
Umm.

 **Anand sivaraman** 27:41  
Our our roadways updated.  
Can you explain the question a little bit, father?

 **Speaker 1** 27:45  
So I'm I'm I'm thinking and I think doing this is a little on the data model side rather than the the vision Kafka topics setup user because I was thinking how to set them up.

 **Tuhin** 27:51  
Yep.

 **Speaker 1** 27:58  
So I'm because I see in that sheet when we went through that a lot of although it's one table that we see right, but a lot of records, not only records, it's just one particular part of the record is updated based on some table updated, right?

 **Tuhin** 27:58  
Umm.

 **Speaker 1** 28:16  
So that I feel but one follow not the entire record.

 **Tuhin** 28:16  
Correct.

 **Speaker 1** 28:22  
I mean, I'm meaning the only one cell, right?  
One data point that is to be updated so.

 **Tuhin** 28:27  
Hmm.

 **Speaker 1** 28:31  
That is what I'm thinking is that something that will give us the performance or it has because it has been done earlier as that everything was calculated and in there or it was calculated on the way when the report is called so you can at least see all of them, right?

 **Tuhin** 28:45  
Hmm.  
Yeah, I think.  
I'll, I'll.  
I'll try to take a stab at it then such and you may have to jump in and and add it.

 **Speaker 1** 28:56  
Umm.

 **Tuhin** 28:59  
I think the thought process here was that you are correct, right?  
Because and to to answer your question, how frequently those cells, let's just call it ready because you you mentioned the word cell, those cells would be updated.  
It would be quite frequently right?  
Maybe not for all of the rows, but across many rows we are going to see that change because essentially what it is as as the records hits our transaction table, which is kind of like the main entry point right to the system, all of the things that you saw like you know, available point points expiration, et cetera, so and so forth, right, uh, all of like there is going to be some impact if as long as there is a positive or a negative points outcome, right, no.

 **Speaker 1** 29:20  
Umm.

 **Tuhin** 29:50  
No, dad.  
Being in mind your second question was that was your second question that uh is doing those updates more efficient?  
Is it from kind of a the ETL or like you know the the Afghan messaging processing point of view are you are you saying whether that's efficient from a read point of view from the application just want to make sure that I understood correctly.

 **Speaker 1** 30:20  
So not for not on the read point, because we'll have it all done there, right?

 **Tuhin** 30:22  
Priyanka. Yeah.  
Yeah, sure, yeah.

 **Speaker 1** 30:26  
So what I'm thinking is the entire tables that the the what I'm calling problem is like site ID.  
Something if you can bring that up again on the screen, that will be helpful.

 **Tuhin** 30:35  
Sure.

 **Sachin Naikwadi** 30:40  
Yeah.

 **Speaker 1** 30:46  
So if the so this is a single record, right?

 **Sachin Naikwadi** 30:47  
No.

 **Speaker 1** 30:52  
If we look at it as a single record, everything is based on the user ID.

 **Sachin Naikwadi** 30:53  
Yes.

 **Speaker 1** 30:56  
There's a there's two of them, right?  
User ID and UID but if they for example go to total credit points right, these credit points are as we talked about, they are getting accumulated right whether it's deleted or whether it's like deducted or and this is based on the user create.

 **Tuhin** 31:06  
No.  
Alright.

 **Sachin Naikwadi** 31:09  
Yes.  
Yeah.

 **Tuhin** 31:12  
Yeah.

 **Speaker 1** 31:15  
So I think I'm it's a correct me Sachin doing.  
If I'm wrong, this is being like if a credit point comes in right or it's the anything in the order also is deleted, this is updated.

 **Tuhin** 31:24  
Yes, it's it's OK translator credit or a debit transaction, right?

 **Sachin Naikwadi** 31:26  
Yes.  
Yeah.

 **Tuhin** 31:28  
Create, add or debit right?

 **Speaker 1** 31:29  
Umm.

 **Tuhin** 31:30  
So yeah, from that point of view, correct.

 **Speaker 1** 31:31  
Yes.  
So so how many of these I'm thinking?  
And no, there's.  
I'm like you can.  
I've been torrent in this.

 **Tuhin** 31:41  
Yeah.

 **Speaker 1** 31:41  
Can you report access to main? Right?  
So how many of these the the columns that you're calling right?  
The calculated ones that are going to be impacted when we calculate this and where is it coming from?  
Is it just a single table that's updating that or is it like we are?

 **Sachin Naikwadi** 31:56  
Ohh Priyanka.  
The thought process was umm, so initially we will fill up this OK, so if user perform any action, let's say he earned 10 points.

 **Speaker 1** 32:06  
Umm.  
Umm.

 **Sachin Naikwadi** 32:11  
OK, so we have to update this field that.

 **Speaker 1** 32:11  
E.

 **Sachin Naikwadi** 32:17  
Plus whatever points come into this.

 **Speaker 1** 32:21  
You, yes.

 **Sachin Naikwadi** 32:22  
So we're going to not again take the summer, yeah.

 **Tuhin** 32:27  
Uh, yeah, I, yeah.

 **Speaker 1** 32:28  
So it.

 **Tuhin** 32:29  
And to answer your question, before you move on, you had another question.

 **Speaker 1** 32:31  
Umm hmm.

 **Tuhin** 32:32  
Priyanka, right.  
So based on what type of transaction is here, right?

 **Speaker 1** 32:39  
Umm.

 **Tuhin** 32:41  
You could have one column updated or you could have more than one column updated, right?  
So actually let me take a look at it.

 **Speaker 1** 32:49  
Anand.

 **Tuhin** 32:50  
So at bare minimum you're gonna have total credit, total debit or redeem or expired out of this four, right, uh for one of them is gonna be updated right searching out of those photo of them then always the balance is going to be updated.  
Right.  
So at least you're gonna have two plus in any scenario.  
And in terms of source such in its coming from just the transaction table correct this this changes so it's coming from a single table.

 **Sachin Naikwadi** 33:22  
Yeah.

 **Speaker 1** 33:26  
OK.  
Hey, I think there's that make sense.  
So if it's coming from the transaction table then I we can track it using the user ID ID has any transaction changes right?  
That is a new input in our transaction table itself, right?  
So it's not getting from anywhere else but the trunk, and I'm going, right?

 **Sachin Naikwadi** 33:52  
Yeah, because user ID is there in the queue for any incident happen.

 **Speaker 1** 33:56  
Umm.

 **Sachin Naikwadi** 33:58  
It happened on transaction table so we can do come user ID.

 **Speaker 1** 34:04  
Like so, it will. No.  
Why?  
I was thinking is the transaction tables because it's coming from the transaction into the detail table or you know the additional but a transaction table right?  
So it creates a record whenever there is a new, like uh, via a buying order and you have a credit in there.  
So that's one table and you get everything there and I I do have that.  
I was trying to look at the transaction table there but anyways so I think yeah, we can look at that and we can go through that because you are here, you are using user update data in one table.  
We are looking at it as one table, but they're in the my sequel where they're getting from right wing one where the transaction actually happens.  
There'll be a record created on.  
Based off that, we see if there's any changes and we'll calculate that will become our topic and we can put it in Lambda or S3 and then calculate it.  
And my understanding correct Sachin.

 **Sachin Naikwadi** 35:04  
Yeah.  
So in Lambda we have to figure out if these tables are get updated.

 **Speaker 1** 35:11  
In.

 **Sachin Naikwadi** 35:12  
User and if that table in that vein log then we have to perform a operation of this.

 **Speaker 1** 35:21  
OK, OK.  
I think umm, I know that that that answers my question, OK.

 **Jubaed Prince** 35:31  
Umm.

 **Speaker 1** 35:31  
The main thing will be go ahead and.

 **Jubaed Prince** 35:35  
Yeah, go ahead.  
Finish that talk.

 **Speaker 1** 35:37  
You know, I was thinking that main now we will for these I think table success is not a problem.  
The main will be there.  
Are we go?  
What topics are going to update them where the update is and then going to calculate right?  
It's not that topic is not going to calculate, but it's going to update us the for example, the credit point line tire record, then when update comes in and that has to be calculated either directly on the fly in the Lambda function or then we put it in S3, yes, we get pick it up based on like OK, there's a new record coming in and then you want to jump you if I'm going.

 **Anand sivaraman** 36:14  
I I I agree.  
So I agree, Priyanka.  
So typically what happens is when we deal with streaming data the the aggregations and calculations umm being performed can have an impact on the on the streaming data.  
So unlike Kafka, if you think about tools like Confluent, which is basically built on top of Kafka, confluent also gives you what is called as Kafka SQL.  
I think this is also there in Kafka but they give you Kafka C SQL that can be used for all kinds of etls during the streaming process.  
But to to Priyanka's point that we just have to be a little careful in terms of adding more aggregations during the streaming process so.  
Affect the overall timing of updates, right?  
This was this has been some experience that we've had on that before, but we can always explore that for them.

 **Tuhin** 37:16  
So if we on that thread just a little bit farther Anand right hippotherapy if we don't put that aggregation there right?  
This aggregation at the end of the day would have to happen at some point, right?  
What would be an alternate alternate approach there?

 **Anand sivaraman** 37:41  
Alternate could be UM as and when, yeah.

 **Tuhin** 37:42  
Is it?  
Yeah.  
Sorry, go ahead. Yeah.

 **Anand sivaraman** 37:50  
Don't know what we can do is we can group the.  
But which is actually the thought process of bringing in streaming?  
If I understand it right, we can group the aggregations together through a faster process which is in the background and then and then updated.  
That is something that we can explore, else we need to.  
We need to think about bringing in some kind of a, you know, streaming SQL that comes with Kafka.  
I'm not sure if the current setup we have is using case SQL, else we can explore case SQL as part of it because that is brought in as a that is brought in as a as a tool for handling this beta.

 **Tuhin** 38:20  
Umm.  
Yep.

 **Anand sivaraman** 38:34  
These kind of requirements better.

 **Tuhin** 38:38  
And are you kind of thinking about uh Anand?

 **Anand sivaraman** 38:40  
And then.

 **Tuhin** 38:44  
I guess what I'm getting to work. Yeah.

 **Anand sivaraman** 38:50  
So you and I, I love.  
I lost you for a few seconds.

 **Tuhin** 38:53  
Can you hear me?

 **Anand sivaraman** 38:53  
Sorry. Yeah.

 **Tuhin** 38:54  
Give me.  
Give me one second.  
I think I'm having network.

 **Anand sivaraman** 38:56  
Now yes.

 **Tuhin** 38:56  
You should give me, yeah.  
Can you hear me still Anand?

 **Anand sivaraman** 39:03  
Yes. Yes, Sir.  
Yes, Sir.

 **Tuhin** 39:05  
Hello.  
OK, OK.  
So what I'm getting towards, and this may be a layman's understanding, right?

 **Anand sivaraman** 39:07  
Yeah, yeah.

 **Tuhin** 39:12  
So let's actually hypothetically do this.  
Who won?  
Exercise really quickly, which I was just being up something in such a Neha, you'll have to correct me if this numbers that I'm going to show are even in a ballpark data correct or not.

 **Sachin Naikwadi** 39:38  
Yeah.

 **Tuhin** 39:41  
So one second.  
Sorry to take on a detour because I'm just really making sure that I'm understanding this.  
So can you guys see a Google spreadsheet?

 **Speaker 1** 40:09  
Yeah.

 **Tuhin** 40:10  
OK, Sachin, uh, I think we have about 19 million record.  
Last time we checked on a clean basis across AP, US and EU like on the transaction table.  
Uh, maybe.  
Maybe like you know, those are essentially the by the kid.  
They if I do an aggregation that's kind of the value I got last time, does that kind of sound sound about accurate that 19 million?  
Just net new transaction if you have to guess ballpark.

 **Sachin Naikwadi** 40:48  
Yeah, yeah, yes.

 **Tuhin** 40:50  
Yeah.  
So I essentially said instead of 19 like I said 50, but you know, just to overestimate this as well as you know, maybe Sachin, we are doing some updates that as well, right, et cetera whatever like let's just call it 52 data center, right?  
And if I just take the US as an example, let's call USA is 25 million network, right?  
That is monthly, right?  
Just for taking 50% of this value, if we look at a daily value.  
And again, I'm.  
I know I'm talking about the mean, and I'm not talking about the spikes, right, but it 100,000 decodes a day an hourly basis.  
It's probably going to be.  
Had 4000 right and we're probably talking about it somewhere in the neighborhood of 600 record per second.  
If it was an uniform right, but we know that distribution is not going to be that way, right?  
So the reason I'm bringing this up Anand right now, coming back to the question right, you know is that.  
Is that kind of seems manageable from a streaming point of view, right?  
If we're going to create like a application on top of the streaming messages that we are getting from Kafka or is that, do you think obviously like let's even assume the higher number per second instead of doing 500 and like 1000 record, let's just call it right?  
It's an app hour.  
We we're gonna take the hours volume within a second, right?  
I don't know if that's a good framework to think about it based on that data points that I have, but I just want to get your thoughts right.  
If you're talking about volumes around that right, you know, do you think we should really think about, think about a specialized tool or do you think it's something that we can try to handle?  
I don't know with with other technologies like it's it it's apathetically question but I wanted to put some numbers around.

 **Anand sivaraman** 43:01  
OK.  
Umm, good in my humble opinion Tuhin the this number that you're seeing right?

 **Tuhin** 43:07  
Yeah.

 **Anand sivaraman** 43:08  
579 per second can can be handled outside a streaming approach also right?

 **Tuhin** 43:09  
Yeah.

 **Anand sivaraman** 43:16  
But but it also depends on what is the rate at which you are proliferating the volume of data.

 **Tuhin** 43:16  
Umm.

 **Anand sivaraman** 43:21  
So today it might be 579 tomorrow.

 **Tuhin** 43:22  
Correct.

 **Anand sivaraman** 43:24  
If it is like a, you know you are, you are exponentially increasing.  
That's when streaming tool like Kafka can really play a better, better role.

 **Tuhin** 43:28  
Umm.

 **Anand sivaraman** 43:33  
So my my so my question is not about using Kafka.  
I that is a strategy involved.  
My question is if we are using Kafka, do we have to include all the aggregations and data manipulation in the same stream or do we bring in do we use Kafka as an ingestion tool only and do the aggregation and data manipulation in a so that was what I was I was trying to understand in the whole process.

 **Tuhin** 43:43  
Umm.  
Umm.  
OK.  
OK, OK.  
So maybe we take a look at it from the picture point of view here.  
Uh then, Anand bread.

 **Jubaed Prince** 44:12  
Ohh so I wanna give her time check.  
We have like I I think we should do this part of the exercise for 15 more minutes and then jump into other topics for the day just to make sure we're being efficient with that time. Sorry.

 **Speaker 1** 44:36  
That was done too.

 **Anand sivaraman** 44:38  
Surprise. Sure.

 **Tuhin** 44:38  
OK.  
Yeah.  
So in this case, Anand, right, So what are you saying that, OK, so the streaming snappening right, are you essentially talking about this approach here, right?  
Storing that data, then doing the manipulation afterwards is that the distinction are we talking about or are we?  
Are we talking about the difference between option A and option BI guess?

 **Anand sivaraman** 45:06  
No, I'm.  
I'm I'm talking about uh, at least maybe Priyanka.  
Correct me if I'm wrong or I heard it wrong.

 **Tuhin** 45:09  
OK.

 **Anand sivaraman** 45:11  
I thought during the streaming itself there were some aggregations that were done.  
There were some calculations that were done right as soon as as part of the topics itself frame.  
If I heard you right in the topic itself in the Kafka topic itself, there were calculations happening right?

 **Speaker 1** 45:22  
Umm.

 **Anand sivaraman** 45:26  
If I heard you heard you right.

 **Speaker 1** 45:29  
Yeah, actually Anand idly, yes, thinking like the do we do it and after topic or do we do it in Lambda function and then processes my concern here is that it's going to be like we have 576.  
Let's go per second right 576 every update that comes actually calculates right file and sent right?

 **Tuhin** 45:47  
Yeah.  
Umm.

 **Speaker 1** 45:52  
So that calculation you have to do it on the fly if you want it real time right Anand.

 **Tuhin** 45:53  
Correct. Correct.

 **Speaker 1** 45:58  
That's what you're also talking about.

 **Tuhin** 45:58  
Umm.

 **Speaker 1** 46:00  
And is it?  
I'll ask a separate question to in I'm wondering, is that real time every single update calculation is very, very important rather and because that is a cost concern, right?  
Or is it like a minute within a minute?  
If we run the function, it can calculate everything we have to make sure that within a minute every calculation is.  
Update we do it in a minute.  
You get where I'm coming from because that actually puts your cost point on a very fixed one, rather than running every update and running division.

 **Tuhin** 46:37  
Umm.

 **Speaker 1** 46:40  
Kafka and Kafka is just a messenger, right?  
We can calculate that and as Anand you earlier said, we can have confluent Kafka or we can have the calculations like Kafka is the consumer is the Lambda and it's calculating.

 **Tuhin** 46:44  
Umm.

 **Speaker 1** 46:55  
Am I right on new?  
What do you think?

 **Anand sivaraman** 46:57  
Stop shop.  
Shop maybe the the the point that tune just to answer your question.

 **Speaker 1** 47:02  
Umm.

 **Anand sivaraman** 47:03  
So let's go by this approach, right?  
So we have debezium to to handle the change data capture that is then being pushed to Kafka for all the real time streaming and then we have a option B which is a fail safe approach where we are actually placing all the data in S3 buckets so that in case anything fails we can again fetch the data from history right?

 **Tuhin** 47:06  
Umm.  
Umm.  
Umm.

 **Anand sivaraman** 47:25  
So that's a that's the.

 **Tuhin** 47:26  
Correct. Yeah.

 **Anand sivaraman** 47:27  
So that through a retry mechanism which which I E of which I think is a is a good approach at least to start with.  
On option A, if I if I understand it right option A, we are using Lambda to fetch the topics from Kafka and then using Lambda we are updating my sequel.

 **Tuhin** 47:41  
Hmm.

 **Speaker 1** 47:44  
Umm.

 **Anand sivaraman** 47:45  
So I don't see anything.

 **Tuhin** 47:46  
Kind of.

 **Anand sivaraman** 47:46  
I don't see anything wrong with this approach.  
Option.  
Yeah.  
The only the my apologies if I missed something.  
The only point is Kafka should be brought in or used more as a streaming tool to ingest data.  
You add more and more weight, which on calculations inside Kafka inside the topic.

 **Speaker 1** 48:01  
Umm.

 **Anand sivaraman** 48:05  
Then you might you.  
You are typical streaming approach can become overloaded and Kafka is also you have to start looking at maintaining Kafka topics and clusters and you know auto scaling of those clusters.

 **Tuhin** 48:17  
Umm yeah.

 **Anand sivaraman** 48:18  
All of that will will be a much challenge, so as long as we are able to do the aggregations and calculations in either, let's say serverless functions using node, you know that would be the ideal way.

 **Tuhin** 48:26  
Umm.

 **Anand sivaraman** 48:30  
That's all I'm trying to say.  
Finitely one option which I would go for.

 **Tuhin** 48:34  
The same page then right? Because.

 **Anand sivaraman** 48:37  
Sorry dude.

 **Tuhin** 48:39  
No, no.  
Good, good. Yeah.

 **Anand sivaraman** 48:41  
To option B, the fail safe approach is something that I would go for, even if it takes a slightly additional time, primarily because that gives us an opportunity to even if let's say something fails, we get an opportunity to pick up the data from S3 and appropriately a continue the pipeline, right?  
That's that's I would, I would say so the there is 1/3 option to in what the third option is let's say we are not using serverless function by using what I was mentioning was by using something called as Kafka Sequel or Case SQL.  
All the necessary weight lift, you know, heavy lifting on terms of transactions or aggregations can be done within the Kafka topic itself if need be, but that would that is something that is needed only when we are looking at large volumes of data and on the fly ETL during the streaming itself.  
But in our case with 579 records and even if it proliferates month on month at 10% either option A or option B both look OK Steven.

 **Tuhin** 49:36  
Umm.  
And in terms of if I understood it currently correctly, the options you kind of sit here essentially within the?

 **Anand sivaraman** 49:48  
Yeah. Sorry.  
Yes, yes.

 **Tuhin** 49:51  
Holding the topic is it's OK.

 **Anand sivaraman** 49:53  
Yeah, but that comes with the costume that comes with the cost.

 **Tuhin** 49:56  
Sure, sure, sure.

 **Anand sivaraman** 49:56  
And if we are not able to use use the streaming process properly, it can add you know it can add a little bit of it again can lead to our data bricks kind of environment if we are not utilizing that properly.

 **Tuhin** 50:09  
Sure.

 **Anand sivaraman** 50:09  
So uh option here and option B are time tested.

 **Tuhin** 50:10  
Yeah, yeah.

 **Anand sivaraman** 50:14  
Tune right?

 **Tuhin** 50:15  
OK, OK, sounds good.

 **Anand sivaraman** 50:20  
So earlier just wanted to be sure.

 **Tuhin** 50:20  
Makes sense.  
Uh, no.  
No, no, no, no, no, no.  
Make it makes sense.  
So in that case then I think Priyanka coming back to your point about like you said like do you need to have it real time to the second or do you have a minute, do you have opportunities for like a minute umm A to do the, can you talk through those uh a trade off please on that in in terms of you know I think I get it on a high level but wanted to hear from you right?

 **Speaker 1** 50:42  
Umm.  
Umm.

 **Tuhin** 50:55  
So like you know, let's just talk through.

 **Speaker 1** 50:55  
Chuck, I think umm.

 **Tuhin** 50:57  
Yeah, yeah, yeah.

 **Speaker 1** 51:00  
I think they're the same points as Anand just mentioned because I was also thinking on the same thing.  
For example, in the Kafka right?  
Then we're thinking about right now is that they're only thinking about ingestion is using debezium coca, and then you know directly then we have a serverless function.  
Let's take for now Lambda function or not just Lambda function consumes from classical calculus and then put it into your reporting DB, right?  
But every second you have 597, so you have to also, you know, make sure that you have to scale the Kafka right to do that rather than that actually a better option is what online said, we can actually calculate in the topic itself, but it puts a lot of load right for each of your read.

 **Tuhin** 51:24  
Umm.

 **Speaker 1** 51:41  
It will also calculate it's not going to, it's going to see through everything because there's very, very Data right coming from, it's not like 1 user.

 **Tuhin** 51:44  
Yep.  
Umm.

 **Speaker 1** 51:51  
Everything is done and we put it there, right?  
So that is what is making me.  
After having that much fun is time.  
Another is cost that is going to, you know, it will always be up and it will always be done it.  
Then we'll again be at the data bricks level, right?  
The other option the be the fail safe right that is.  
There you are running it.  
It's ingesting.  
It's simple.  
Kafka running and ingesting and putting 597 records in the specific files the same file so it's quicker but then you have the Lambda function and this is a fail safe.  
Also that you always have the data need it right in your in an St bucket and then your Lambda function.

 **Tuhin** 52:33  
Yeah.

 **Speaker 1** 52:34  
What I'm thinking is every minute we can run that and it can set your data calculated and put it in my SQL.  
I mean, do you think I'm going in the right direction with the cost?  
Because I'm right now, it's not the processing processing is there.

 **Anand sivaraman** 52:47  
I.

 **Speaker 1** 52:48  
If you put it in the Kafka, but I'm concerned about the processing, it will take more time and still cost and I'm also concerned about the cost in this scenario in the real time ingestion.

 **Anand sivaraman** 53:01  
I I I see the point, Priyanka.  
But I think rather than discussing this, it would probably be easy for us to very quickly do a small POC and then test this out if my my.

 **Speaker 1** 53:04  
Umm.  
Hey.

 **Anand sivaraman** 53:14  
My point is there is a.  
There is a reason for bringing in streaming here, and that that streaming pipeline should only be used for streaming, not for, not for all the calculation, Analytics & aggregations and business rules are going to go or only going to increase as we bring in new and new and new data if or as the business changes.

 **Speaker 1** 53:26  
Yes.

 **Anand sivaraman** 53:35  
So all the that that there needs to be proper bifurcation between what is being streamed, what is being rangled or processed in terms of the in terms of the data transformations and what is being stored.  
So from that standpoint, option A where Kafka, serverless function and MySQL and option B where Kafka you know save it in S3 and then move it to Lambda and my sequel are time tested approaches.  
If we are really looking at faster updates into my SQL without using Lambda or without using S3.  
I I stand and I I just repeated what I said some time back as well.  
So I I would, I would say going with the time tested approach of option here, option B seems to be a better option.  
We just need to be sure that in the don't bring in all additional.  
Transformations that are needed for the streaming purpose.

 **Jubaed Prince** 54:37  
Umm I have a I have one question here which is is it possible to view the five points my SQL five point Sevens is that that runs on a MySQL 5.7 DB.

 **Tuhin** 54:38  
OK.

 **Anand sivaraman** 54:42  
Yes, Sir.

 **Speaker 1** 55:00  
The things I can you repeat the question?

 **Jubaed Prince** 55:04  
So let's say that our my school 5.7 D with their current the production DB that we have, it has SQL queries running right all the time.  
So is is there any way to know how many queries and what were the queries that were running last one hour?  
From my SQL.

 **Speaker 1** 55:35  
I'll.

 **Sachin Naikwadi** 55:35  
Yeah.

 **Speaker 1** 55:38  
You can't.

 **Sachin Naikwadi** 55:39  
You mean today?

 **Speaker 1** 55:39  
Once we can do that, yes, it's but the.

 **Sachin Naikwadi** 55:39  
Tonight you run running beauty.

 **Karthik Raghu** 55:39  
What is?  
What?  
What is the intent, Prince like I understood the question, but I did not.

 **Jubaed Prince** 55:49  
So the intent is if if you copy everything and put it on real time on a SQL, that's basically what tracking all that changes in a S3 right?  
So if I have that changes already somehow logged in my SQL and if I can get that, then I don't need all these middlewares, I can just copy them and process them one by one.

 **Karthik Raghu** 56:15  
No, that's, that's where you need the middleware, right?  
Like it does log all this data in in the bin log.  
So the bin log is is where.

 **Jubaed Prince** 56:22  
So been love does that.  
Is there any activity log or any sort of uh log that that tells us how?  
What were the queries that were ran on that within a particular time?

 **Karthik Raghu** 56:41  
Umm.

 **Jubaed Prince** 56:44  
So I can see there is like a query I I'm just reading some random.  
Posts but it says there is something called general log that you can turn on to see my SQL query log.  
Can we utilize something of that nature?

 **Karthik Raghu** 56:59  
No activity.  
Everyday will log will be there right?

 **Muniaraj Thangavelu** 57:03  
Do you?  
The actually ohh it is user based I the client based on different scenario.  
Different update will happen right?  
So everything calculate and based on the query log it is bit complex and cumbersome because maybe sometimes the user table updated with particular user and again point is updated by the transaction.  
Everything will be difficult, right?

 **Jubaed Prince** 57:31  
Yeah, let's talk about the difficulty a little bit later, but is it possible to get the log of the queries?

 **Speaker 1** 57:36  
Yes, uh, Princess.  
Yeah, it is possible, but it will also impact the performance of your server.  
You can enable the logging that you mentioned right?

 **Jubaed Prince** 57:50  
To a POC around that as well.

 **Speaker 1** 57:50  
Because it's always using.

 **Anand sivaraman** 57:51  
But but Priyanka.  
Yeah, yeah, sure.  
Prince Priyanka can be I I I remember a cloud insights capturing my SQL activity as well, particularly if it is part of Aurora RDS.

 **Speaker 1** 58:03  
Hey.

 **Anand sivaraman** 58:05  
So can we can probably like print the suggesting we can do a quick POC and have the team work on that, yeah.

 **Speaker 1** 58:13  
But that I I'm thinking the version that that that was the same thing I was thinking, but I'm thinking is 5.7 and like can be written that on that's the it's still version AWS still version that impacts it impacts the queue.

 **Anand sivaraman** 58:31  
Hmm.  
OK.

 **Karthik Raghu** 58:31  
I what as I know I'll tell you.

 **Anand sivaraman** 58:32  
OK, that's a good point. OK.

 **Karthik Raghu** 58:34  
I might be wrong, but as far as I know I remember activity log contains information about it.  
It actually tells you what what happened.  
I mean like it's just an activity, right?  
Like let's say the user is deleted, it just says a user is deleted.  
Umm.  
And then let's say a user is added. Uh.

 **Jubaed Prince** 58:53  
Does it say in the SQL format?

 **Karthik Raghu** 58:58  
No, no, that that's what right, it's it's an.

 **Jubaed Prince** 59:00  
How does it say that?

 **Karthik Raghu** 59:02  
So it logs the activity like it just gives you tells you information about what was done.  
What happened?  
I mean like it, it's more of more on the log standpoint, right?  
Like it's it's more human readable, so this is what we usually ingest in in Splunk kind of tools to identify what activity happened.  
Let's say like you enable audit logs right?  
Like you enable audit logs, you enable activity logs.  
So mostly what what happens is like, you know, as far as I know it, it's more on the human readable standpoint.  
So that's why I think you know ohm.  
That's why we have Bin Log which will actually have all the information log along with the queries right?  
So it will log the queries and then based on those queries.  
Uh, we will be able to that.  
That actually opens up, uh, possibilities of actually, you know.  
Uh.  
Having the ability to restore the database to any point in time.  
Uh, because we have all those queries, right?  
Uh, as far as I know, I think Bin Log is is where we get all this query information.

 **Tuhin** 1:00:14  
So, Prince, can you help me understand this?

 **Jubaed Prince** 1:00:14  
So.

 **Tuhin** 1:00:16  
Like, let's assume that Corey is existed.

 **Jubaed Prince** 1:00:17  
Yeah.

 **Tuhin** 1:00:18  
All of the insert, post, deletes whatever right updates existed and you could get it.

 **Jubaed Prince** 1:00:21  
Hmm.  
Hmm.  
Umm hmm.

 **Tuhin** 1:00:24  
What did you do with those queries?

 **Jubaed Prince** 1:00:27  
So then I will not stream it.  
I will.  
Will, as Priyanka was suggesting, I will check it every minute, every minute.

 **Tuhin** 1:00:33  
In run the same queries on my circulate.

 **Jubaed Prince** 1:00:38  
So every minute of function will.

 **Tuhin** 1:00:38  
So you check it, you got the coding, yeah.

 **Speaker 1** 1:00:42  
So.

 **Jubaed Prince** 1:00:42  
And and then I'll.  
I'll just do the needful in my escalate.

 **Speaker 1** 1:00:48  
So, Prince, actually that is.

 **Muniaraj Thangavelu** 1:00:48  
Cheap actually.

 **Speaker 1** 1:00:52  
One second.

 **Muniaraj Thangavelu** 1:00:52  
Yeah, go ahead.

 **Speaker 1** 1:00:53  
So Prince, the purpose that you're saying that is now in my SQL, right?  
You can enable the CDC which reads which actually starts capturing your log every time there's an update.  
There's a log and I think that's where Karthik bin log is.  
Also bin log that Karthik is mentioning that.

 **Jubaed Prince** 1:01:07  
No, I get that bin log we have been discussing, right?  
I'm I'm asking if we if you are doing a minute by minute processing then you don't need to feed in the data every minute and store it in SS.

 **Speaker 1** 1:01:10  
Yeah.  
Umm.  
Exact actually.

 **Jubaed Prince** 1:01:20  
Essentially, you're doing the same thing.

 **Speaker 1** 1:01:21  
Yes, exactly.  
And that is so that is what we were alluding to earlier, right when we started the conversation, if when, minute by minute is something that we can look at, then it'll the same thing right now you can have it in S3 format and the Lambda function can look at it and let me look at the Lambda function can also look minute by minute into your.  
A buddy lock is a little by yes, so there are two types of log.

 **Jubaed Prince** 1:01:46  
Into that logs that my SQL has right cause if you are copying the data to SQL Yeah 2.

 **Speaker 1** 1:01:53  
Hmm.  
Here's a general query log that.

 **Jubaed Prince** 1:01:57  
Log handy or something? Yeah.

 **Speaker 1** 1:02:01  
So that that'll give you very simple format SQL query.  
What you are asking for, right?  
But it is not very effective in the when any function on any read will take that.  
That'll look at the bin log because that is in binary format, so it's faster.  
That's why Bin Log is used for reading a I think the having what kind of log and if we are OK to read every minute right from that bin log is a different think.  
If I'm understanding that experience, you are also going on that file that can be read directly from that change log or query log what query has been done and update or anything on.  
Then calculate.  
Am I correct?

 **Jubaed Prince** 1:02:43  
Box, yes.

 **Speaker 1** 1:02:45  
So this actually comes back, yeah.

 **Jubaed Prince** 1:02:46  
Has that could potentially so that the thing I'm trying to get at is.

 **Speaker 1** 1:02:53  
Umm.

 **Jubaed Prince** 1:02:53  
I like the fact option of running it every minute.  
We maybe we have a like a 62nd delay on the reports, but.  
And and we can probably decrease that time as well, right?  
We can run it per second as well, but the question is if we go copy the bin log items real time are in an S3 versus can we just directly go to my SQL L 5.7 and just get the log and somehow get that information and run it in every minute or every second?  
So instead of running that real time streaming, can we do something else around that with the existing data?  
Because eventually what you are necessarily doing is copying it and keeping it.

 **Speaker 1** 1:03:40  
So.  
So. Umm.

 **Jubaed Prince** 1:03:44  
There's three and doing the same thing, right?  
So I'm just trying to avoid that streaming if that's possible.

 **Speaker 1** 1:03:51  
So actually, Prince uh, let let me put it that way that then you you are talking about reading right it you're saying updates right.

 **Muniaraj Thangavelu** 1:03:52  
Stupid the action.

 **Speaker 1** 1:04:02  
But we also have to calculate this is 2 steps that you have right.  
One is as it is, whatever is coming that is not coming as calculated for your report.  
So yes, we can read that and that even on the lower side, if it's only a read, that is what I think earlier and was also saying that if it's a read you use debezium you use Kafka and it reads wherever you want to read anything can be a sync, right?  
That does not take time, just direct ingestion.  
OK.  
So that is where you you don't even have to go for a minute.  
You can just read real time.  
That is alright if you so my what?  
My concern, and I think Anand is also talking about, is the calculations that we have to do further for your reporting there.  
Actually, this is exactly the same thing I'm saying.  
We can actually read or we can calculate every minute, but your files will be there, right?  
There is a CDC update in your file system and Lambda every minute can read from there.  
If you want the Lambda to read from this thing, not even Lambda.  
If you want to directly then he is again going back to where your system is right now.  
Reading the SQL query and then calculating it so you have to do it somewhere right?  
They hope uh like.

 **Jubaed Prince** 1:05:20  
So just help me understand this part.

 **Speaker 1** 1:05:22  
I'm making sense. Umm.

 **Jubaed Prince** 1:05:23  
Let's say you don't have cafka.  
OK, so now you have a bin log enabled.

 **Speaker 1** 1:05:27  
Umm.

 **Jubaed Prince** 1:05:31  
You have audit log enabled.  
You have my SQL login about whatever logs you have, you have those enabled now.  
Can you use your Lambda to read those changes from the database and make the new changes in the report DB?

 **Muniaraj Thangavelu** 1:05:49  
Stupid.  
Just I want our own point because whatever you are taking log is based on the base table, not in the summary table, so that will not be it's.

 **Speaker 1** 1:05:50  
Uh.

 **Muniaraj Thangavelu** 1:05:59  
If you want table again update to happen in the reporting DB instead of Kafka we can go for mirroring, replicate the replication DB such a way that we need not to do the action again on the DB right?  
So one hour and the 579 is only for the points report our entire my application will be only ever 579 and find 79 will not happen for all the table in a single time.  
See the user email ID will not change frequently, but that received some user table.  
My credit point, whenever the transaction happens, so multiple tables involved to update in this same means won't pay.  
In such a way that.

 **Jubaed Prince** 1:06:48  
But the changes are gonna happen.

 **Speaker 1** 1:06:50  
Umm.

 **Jubaed Prince** 1:06:51  
So there are two steps, right?  
One is reading what happened in 5.7 and then second step is doing something on 8 right?

 **Speaker 1** 1:06:54  
Umm yes.  
Yes.

 **Jubaed Prince** 1:07:03  
So we are talking about reading 5.7.  
First, let's forget what happens in 8).

 **Speaker 1** 1:07:06  
Yeah.  
No, no.  
Yeah.

 **Jubaed Prince** 1:07:09  
So in this reading reading site, yeah.

 **Speaker 1** 1:07:09  
Yeah.  
Let me answer your question.  
Umm.  
Umm, so in right you don't need to read from the bin log which is faster, right?  
It does not directly read those SQL queries.  
Even if you enable and Lambda won't.  
And I I don't keep me honest here, because that is my limited knowledge.  
But you will have to have a tool that will read the binary log.  
So even if you want to do Lambda, you have to either have David setup or you can even have AWS tennis read that.  
But there has to be a tool which reads a bin log on an missing something here.

 **Jubaed Prince** 1:07:47  
But it has to be streamed.  
It has to be streamed.  
It cannot be.

 **Speaker 1** 1:07:51  
No, it does not.

 **Jubaed Prince** 1:07:51  
Just go to the database and check it.

 **Speaker 1** 1:07:55  
For the batch processing, right and so you can do that is what we have been doing right.  
We have been from the SQL Pyspark right when you have he.

 **Jubaed Prince** 1:08:05  
No, just answer me this.  
Just answer me this, is it possible to read that bin log related change or the new queries that are happening in 5.7 without real time, just collecting it through Kafka?

 **Karthik Raghu** 1:08:26  
And they did that already via POC, right?  
Since the real time processing was not there.

 **Jubaed Prince** 1:08:34  
No, no, no, no, no.  
I think we are getting.  
We are not on the same page here my yeah.

 **Tuhin** 1:08:38  
To so Prince can I can I?  
Can I just add from a layman's perspective, right?

 **Jubaed Prince** 1:08:42  
Yeah.

 **Tuhin** 1:08:45  
Are you saying instead of reading the data, let's just read the query?

 **Jubaed Prince** 1:08:45  
Umm.

 **Tuhin** 1:08:49  
That's that's essentially that's essentially all you are saying.

 **Jubaed Prince** 1:08:49  
Yes, yes, yes.

 **Tuhin** 1:08:52  
Like if if you know the query then you can take the query, transform the query and update the summary tables.

 **Jubaed Prince** 1:08:57  
Yes, yes, yes.

 **Tuhin** 1:08:58  
That's what you're talking about.  
OK.  
And what what's your?

 **Jubaed Prince** 1:09:02  
And let's do it.

 **Tuhin** 1:09:03  
What's your thought process?

 **Jubaed Prince** 1:09:04  
Not real time, but let's do it whenever I want to do it.  
I'll go and check.  
OK.  
In last seven days, what were the queries and I'll do my changes accordingly in last one second, what were the queries?  
I'll go and change my things.  
Is that possible or is that you have to stream things all the time?

 **Karthik Raghu** 1:09:19  
I yeah, I I'll.  
I have one more concern here.  
So so even if you enable the activity logs right so they are based on what do you say database like how many databases you have, do you have to do that at the database level?  
Not that like cluster level.  
Even if you do at the cluster level, you'd have to access it via the data.  
That business and they configure those to read, would like check what is happening on on the four databases I'm talking about the cluster setup, right?  
It's not just one single database here.  
Looking at we there could be multiple readers multiple.  
What do you say?  
Uh.  
Writes happening right?  
So as part of the tactility like you know the failover happens, the primary suggest a secondary.  
Let's say we have 01 today as writer and then we start, you know, information from the logs.  
See what is happening?  
Capture all the trains and changes so that becomes a reader.  
We will not see any update queries happening in that database, so in that case like you know, it gets really difficult and messy for us to be able to handle that ohm.

 **Tuhin** 1:10:39  
Will Prince.  
What's the Karthick like, you know?  
Thank you for explaining that.  
But Prince, what's the hypothesis that you are driving towards rather than reading the data and reading the queries like you know, are you just trying to think about it?

 **Jubaed Prince** 1:10:47  
So.  
I'm just trying to get rid of a S3 and use use that log as our S3.  
Basically, so use that log to get that data and not do all this middle processing.  
Just have a short circuit between them.

 **Tuhin** 1:11:08  
But you still have to update like you'd still need an app.  
Let's just call it an app or that that's Lambda or not get whatever.

 **Karthik Raghu** 1:11:14  
Umm.

 **Tuhin** 1:11:16  
It's still need something to read it and update the table, right?  
Forget about the S3 bucket.  
Let's just say we just had option one.  
The line right?  
It's still need that, right?  
You're still not getting around that.

 **Jubaed Prince** 1:11:30  
Getting our own.  
What the the app thinking?

 **Tuhin** 1:11:34  
Yeah. Yep.

 **Jubaed Prince** 1:11:34  
App Thinking app is there.  
That's that's that has to stay.  
Another question is how does thinking app gets his data to sync?  
Then there is two options there, from S3 or directly from my SQL.  
That's that's what I'm trying to say.

 **Tuhin** 1:12:02  
It can.  
It can get directly from my skill right from the Kafka topic, correct?

 **Jubaed Prince** 1:12:07  
No, let's just forget about Kafka topic for a bit.  
Just can you do it without Kafka topic?  
Just get it from my school.

 **Speaker 1** 1:12:13  
So no Prince, you know other RDS or still does not allow the straight access of those events to users, right?  
You have to use from bin log.  
It has to be set up so that it can be read right so that that middle person to read that will be there any spark session or anything cannot directly read the window.

 **Jubaed Prince** 1:12:34  
So Bin Log has to stay.  
It has to stream the data.

 **Speaker 1** 1:12:37  
Yes, you can have it in batch processing.

 **Jubaed Prince** 1:12:37  
There is no way to go and see that historical data you get.

 **Speaker 1** 1:12:42  
He yeah, you can do by batch process.

 **Jubaed Prince** 1:12:43  
I can have it in batch processing.

 **Speaker 1** 1:12:46  
Yeah, but you will have to read binary log.  
Then you have to have the tools like debezium.

 **Jubaed Prince** 1:12:49  
Then why if I have?  
If I can do batch processing then why do I need S3 for fails if?

 **Speaker 1** 1:13:04  
OK, let me let me see binary logs that you have.

 **Karthik Raghu** 1:13:07  
No, I'll tell you that the reason why we you know, I tell you the reason why we introduced history, I think yesterday was not there in the picture at all.  
The reason why we added that it's three into picture was like, you know, remember I asked one question like you know I think we need a solution which is scalable and which is also high from the high availability standpoint.  
It should not bring in any new issues with regards to data integrity.  
So what if my cost after server restarts?  
I reboots and use all the data, or maybe maybe I don't know.  
Like you know, whether I come up with a proper uh cluster setup of craft or not, I'm not sure because I don't have that expertise with you, right?  
So what happens if they reboot the craft pass?  
So in those cases I think, uh, I think someone and then you Meeting propose that, OK, if that is the case, I think we can have all those raw data which the into some storage and then we like in since it is like a permanent storage.  
So we will have all the data like even if the server reboots for some reason.  
So it will start capturing the data and then.  
This from this distance.  
Uh, otherwise, I think we never had this idea of drinking in a single storage.

 **Jubaed Prince** 1:14:31  
So it's purely a fail safe purpose, right?

 **Speaker 1** 1:14:35  
Yes, you can directly do it using Lambda also from Kafka.

 **Karthik Raghu** 1:14:37  
Yeah, you you also have.  
It's a fail safe and you also have historical information of what just happened.  
See if you'd like to play around with it in future, or maybe I don't know, like maybe maybe something changes you actually have the raw data.  
Very new, even if the solution changes, the data is seen there like you know, yeah.

 **Jubaed Prince** 1:14:56  
Yeah, I get that cartoon.  
Yes.  
So now if if we can do batch processing, Priyanka how does that batch processing work?  
Can you give me an example?

 **Speaker 1** 1:15:12  
So the Prince actually.

 **Jubaed Prince** 1:15:13  
Instead of real time streaming, how do I do batch processing in coca?

 **Speaker 1** 1:15:20  
You will have to store still you will have to store bin logs right.  
You will have to go through, you know, batch processing is just that it will fetch from the CDC layer and Anand I think this is where I need your input on.  
Augusta of them like something in.

 **Jubaed Prince** 1:15:37  
And study how all this works.  
Right cause it's important how we are accessing the data, how we're storing data.

 **Speaker 1** 1:15:41  
Umm.

 **Jubaed Prince** 1:15:46  
Do we need to make a copy of the data or can we just use the original source of data and use it?

 **Speaker 1** 1:15:47  
Yes.

 **Jubaed Prince** 1:15:51  
If it stores in our MySQL data, then we don't really need a like like another phase safe because it's already in a production database, right?  
It's it's already there, so if we have not tried it, let's try spend some time exploring that part, cause I think that that is important.

 **Speaker 1** 1:16:09  
Umm.

 **Jubaed Prince** 1:16:13  
Why?  
Because if it's three adds a load it it has cost it has.  
You have to build it.  
You have to do all those.  
Can you?  
If you're, if we're doing a minute by minute processing, and if we have all the logs in the MySQL already, then why not just try to use that?  
At least let's give it a try.  
If it doesn't work, then we know, OK, we can go.  
We go with the streaming, but when you say we have batch processing option, my head tells me.  
Inside there and then.  
Then how can we tap that and the?  
The question is, can we avoid this like a backup of our our audit logs?  
Can we just use this this data that this batch batch processing is using?

 **Speaker 1** 1:17:11  
The batch processing Prince I did not mean by batch processing it as in like you have Kafka through binary log and doing batch processing right and the batch processing what we are doing currently in this you have to replicate that CDC somewhere if you want to do batch processing also and that's why we were mentioning Kafka and everything in the middle the batch process.  
The boat set up for your this thing is there.  
What do you call the Databricks right now?  
Right, but data is not doing CDC.  
It's not default setup for CDC right.  
CDC is a concept that you don't want in at least SQL and Anand keep me honest here in in for live streaming that processing you fetch directly from the DV right now, right?

 **Jubaed Prince** 1:17:52  
Uh.

 **Speaker 1** 1:17:58  
And then you ohh put a code in a way that it does change it I captured but that will be on update dates and looking at the logs what was updated earlier that will go but not directly like it has everything in place and you don't have anything even for Kafka.  
One other thing, see this that is in the incremental load which you already have in databricks.  
Right.  
Other thing that it won't deal with this which Kafka and division will deal, you do have to capture them separately is the deletes right?  
Yeah, that have happened.  
If there's no update date right again, go back to where your databricks are raw layer is currently at.  
I didn't I missing something here?

 **Jubaed Prince** 1:18:43  
Why do we go back to Databricks current layer?  
We are having a copy of report DB and aggregating all the reports right?

 **Speaker 1** 1:18:46  
The they're they're not.  
But I'm saying that's what your data bricks layer is doing.

 **Jubaed Prince** 1:18:54  
So it's it's not the current solution.  
Current solution is it aggregating all the results like report DB we were doing here.

 **Speaker 1** 1:19:04  
They're all living.  
Difference is that you can have the direct poll on your.  
The difference I see in the same logic is the you put the incremental you know in not incremental.  
You deal with all these create, update, replace in your spark code and DB data bricks right and you directly put it into the.  
You were there and directly then that is what your batch processing is.  
What I can think of right now for having and and that will take time, right?  
Because it's processing, it's going to read your actual database and it's going to scan this.  
It's not accessing your binary log.  
The only reason why Bin Log comes into the picture is because the processing on the back, the spark can do is read that window right this when we do this on the batch processing directly reading the DB right, the SQL queries that we're talking about that will be your whole table scan or your what updates on the update dates coming in.  
That is what you can read the even if you enable the query log the the umm.

 **Jubaed Prince** 1:20:09  
So if I can read that if I can read that, can I read that anytime I want to?

 **Speaker 1** 1:20:18  
No, you don't about the boundary though.

 **Jubaed Prince** 1:20:18  
Happened.  
From the bin log, yes.

 **Speaker 1** 1:20:21  
So.  
I am not sure because of till now I'm where I'm aware when locks CDC you have to.

 **Jubaed Prince** 1:20:32  
Who works with bin logged here if if this project has to be done, who's gonna implement this one?

 **Speaker 1** 1:20:38  
That we are going to implement that is not a problem.  
I'm saying I bin log is not something that as a user I can directly read right?  
I'll have to read it.  
They have set it up in a way that OK, if something happened and event happened that update.  
For example, if I say it on the AWS side, do you have the cloud bots right?  
Although you have the cloud watch, you can go see.  
You will have to go filter out right?  
That setup has not been done in the.

 **Jubaed Prince** 1:21:01  
No, don't go to Cloud watch.  
Priyanka, you are going to different tools.

 **Speaker 1** 1:21:04  
I'm not.  
I'm I'm comparing, I'm comparing I'm.  
I know I'm comparing cloud watches, something we can read.

 **Jubaed Prince** 1:21:07  
I'm just telling you one simple thing.  
You turn on my.  
You are turning on my SQL bin log right?  
And now Kafka is a separate entity which is consuming the bin log and giving you our real time data.  
I'm asking you can you do something similar and just go ahead and do it every minute?  
Forget Kafka, you have been logged enabled.  
Can you go and see the SQL and do that is is that.  
Is that doable or is that something we can explore?  
Cause if we can, then we don't need an S3 because this data is already there.  
If the data gets hidden, if the data is.  
This solution doesn't work, but do we know if the data stay in my SQL or don't?

 **Speaker 1** 1:21:53  
So. Umm.  
Sure we can.  
We can test that out and come back.  
I think that is something.

 **Jubaed Prince** 1:22:00  
Because if we don't know that, we can go back and learn that.  
But if if we are guessing and trying to just compare solutions that the point is how did Kafka implement it right.  
Kafka is a third party tool, right?

 **Speaker 1** 1:22:13  
Hmm.

 **Jubaed Prince** 1:22:14  
So if they can consume bin log, why can't we now we didn't or that's fine.  
We can explore that, but if we said the Kafka is the only when streaming is the only one, that's a pretty bold statement and we then we need to really go ahead and figure out why is that only like closed like that, right?

 **Speaker 1** 1:22:38  
To.

 **Jubaed Prince** 1:22:42  
So sorry, sorry, but I I think I think there needs to be some experimentation around that and that the reason is if if it works then you don't need a 3 then you can just use bin log and then the step to the thinking part will always happen and it it's it's not the previous solution of course because this is different.

 **Speaker 1** 1:22:43  
I think we can test it out. At least.  
Uh, and see.

 **Jubaed Prince** 1:23:06  
This is happening but this is not happening in real time.  
It will happen every minute and there will be a deal of minute part of the data.  
So that's where I'm like trying to get at because I I believe if we can, if this works, then we don't need S3 and all the expenses all the time to implement that.  
If this doesn't work, if streaming is the only solution, then if we think fail safe is required, then we need an.  
Cool.  
So I want to also jump into the next part of today.  
I think this will help us too, so if you can go to that reporting triage initiative, you will see under March 25 I added a reportable Workshop.  
I added a bunch of our names and please feel free to add your names as well and you can go and check out Sachin's work through from here and then I we are going to ask a lot of questions here.  
Try your best to answer them.  
And I'll, I'll I'll kind of go through each of the question and kind of try to explain where this questions are coming from and feel free to add more questions and more things that really makes us think through these points report.  
So one is.  
This this part is basically the document that Sachin shared in the beginning of today and then for this transform data application DB to reporting DB.  
This Lambda function I I think we can ignore this for the time being, but let's focus on this one.  
What are the inconsistent data points?  
If there are any columns that we are syncing that might change over time, the data can.  
Clean.  
Inconsistent is.  
Is there any sort of vulnerability around that?  
Are there any data point that might cause delay?  
Are there any redundant data point?  
How can we resolve those?  
Is there any conflicting situation for any of the data point and that is for current situation and in future if we try to kind of add more things to it do we conflict anywhere?  
Also, can we think about schema alteration in future where we try to add a new column uh rename a column are are there any problems that we can see in call of when doing those because the computed columns we have a lot of columns that are calculating and also keeping balances and stuff.  
Can we highlight those in here?  
Do we need to lock any row for certain purposes for data consistencies and stuff?  
Are are did we think through that part for each of the data points?  
Are there any challenges during fixing a bug of a computed column in the future?  
So if there is a bug, we find OK point calculation was wrong.  
We are doing it now.  
This way do we think this data model supports those sort of activity for future?  
Is there room for improvement in terms of schema, indexing, stored procedures, partitioning?  
If so, list them down.  
Can we utilize MV and views for for materialized view to just kind of having a cache of the report and views in general?  
Do we need to use them?  
Do we need to think how we are creating, reading, updating and deleting that Data antley?  
So yes, we have a already searching showed how we do it, but can you resync some of the things?  
OK, if you, uh, when creating this keep this data also and then this will help us on that report data point generation things like that.  
Is there any groundbreaking idea that significantly improve the report use case thinking from first principle?  
One of these came, I think, is like that a minute by minute processing that can probably help.  
Is there any other thing that you can think of forgetting all this Kafka and everything?  
Is there anything else you can think of?  
Then you can put that here and what are the data manipulation required in the report DB as the US E One DB changes happen, so where do we need to do this changes?  
OK, I found I insert insert from bin log or CDC.  
Now, what do I need to do?  
What are the changes that I need to do in US E one DB and also I tried to take a note of that discussion solution we had so far Karthik in the message posted.  
One thing about splitting tables based on, I believe customers Priyanka talked about having a regular cadence over having real time.  
We also were exploring real time.  
We have SGS fail safe and unmentioned, KSQL.  
I think this is something we should also explore.  
S3 cash plus per minute sync.  
That's kind of like a combo of what Priyanka was saying.  
And then this is what I'm asking.  
Can we utilize my SQL query log anyway and see how how that those sort of solutions look like?  
If you have any questions, please feel free to ask.  
But other than that, I would really think we should get busy on this dog.  
Put everything in this doc as as we we are thinking doesn't have to be accurate.  
Let's comment on things and just just put everything here so that it's easy to kind of calm down to the final solution.  
I happy to take any questions or improvisation idea in the in the in the process.

 **Anand sivaraman** 1:29:13  
From a Prince, this is a great set of questions.  
Thanks for bringing this up and we'll also add a few more.  
One question is on the there was a point around schema changes right?  
Like as new columns get added, how do we change if I understood that correctly?  
Is that something that you want dynamically without touching the pipeline?  
Is is that the whole thought process behind that question?

 **Jubaed Prince** 1:29:39  
Right now we can since we're making the bees and everything, we can make some notes.  
OK, currently 5.7 has this sort of schema.  
If we do this, that will be better in future.  
Maybe will not touch it.  
That right, it will have it in the list to work on in the future.

 **Anand sivaraman** 1:29:58  
Sure.

 **Jubaed Prince** 1:29:58  
So it's it's it's not ask and and also I think Anand you bring up a good point and I'll add that is alter uh statement in 5.7 how does report do do that right cause.

 **Anand sivaraman** 1:29:58  
Got it.  
Thank you.

 **Jubaed Prince** 1:30:16  
Uh, there's probably so far we're thinking about inside update, delete, but what if the table schema changes?  
How does report DB gets updated based on that?  
Or do we do that manually?  
I think we should also think around that.

 **Anand sivaraman** 1:30:34  
But yeah.

 **Jubaed Prince** 1:30:41  
Cool.  
So I to him, I don't know what?  
What should we do next?  
Should we spend some time updating the doc or should it go as a homework and we do discussions around other things?

 **Tuhin** 1:30:53  
Uh, so?  
I'm happy about all of the discussions, but again just I wrote that down in terms of.  
Uh, you know we can run one experiment to experiment 5 experiments.  
I don't necessarily.  
Uh, I think whatever the team feels comfortable with, but at the end of the day, I need to workable solutions soon because let's keep in mind from the top goal, right?  
We do need to turn off data bricks.  
We do need to get this reports out in life, darling as possible solution.  
So I think we have some good ideas.  
So what I would challenge that team in terms of, but seems viable in terms of expertise and implementation and not discounting any of the concerns that anyone came up with.  
But essentially, how do we go to?  
Time to live.  
Let's call it right, because that's the key factor from the business point of view and that's that's what we have to prioritize.  
So if we want to.  
Child 10 different things we may not have to have the time, so we need to prioritize.  
Which two things that the team is gonna try?  
Because we do need to figure out a solution for each of those concerns as fast as possible.  
Because once this Team kind of bigger figures out the proof of concept, please keep in mind that we would also have to train up the wider team and start implementing rest of the rest of the reports.

 **Speaker 1** 1:32:46  
Doing on the same one.  
Actually I have a question for first which approach do we want to go and look at the bin log or we want to first set up like the vision and see if we can start consuming online.  
Because now we have 3 approaches and Prince you can jump in here, right?  
We want to see if we can consume directly right from.  
The museum cost Kafka into Lambda function calculated and put it into the reporting DB.  
2nd is if we can one part will be consuming it at putting into S3 and the events every minute.  
If I'm not wrong in that that Lambda can run through and update the reporting and the third one is that we directly read from the bin log right and then fetch it into the reporting DB, whatever changes.  
So what do we want to take so that we can start working on that and simply and see?  
Because as you mentioned, the timeline is like Friday, we want to at least have some clarity, right?

 **Tuhin** 1:33:56  
Yeah.  
And I think that's a decision like, you know, I think we should make today.  
Uh, with this team, man.  
That's a that's a purely technical decision.  
I'll leave it up to you the next team and any and the community team, but we don't need to come to that consensus because you know, I EI essentially don't want to end up on Monday and we're still talking through solution, right?  
Uh, all that could be good in an ideal situation where not in an ideal situation right now.

 **Jubaed Prince** 1:34:37  
So there are like 7 things written as solution.

 **Tuhin** 1:34:37  
So how do we figure out which ones?

 **Jubaed Prince** 1:34:41  
I think we can start maybe by commenting on those and kind of eliminating and coming to too.

 **Speaker 1** 1:34:54  
OK, I think the data I will take over and I can see and I maybe Karthik, I can reach out to you when we can explore all these options and quickly close that which one is a preferable one and which two like second option that we can keep open.  
What do you think, Prince?  
Then that will be a good.

 **Tuhin** 1:35:14  
I I think that could be one approach or or we can just talk through it right?

 **Speaker 1** 1:35:18  
Umm.

 **Tuhin** 1:35:19  
Because I think he he he is just now.  
Right, Priyanka.  
Right.

 **Speaker 1** 1:35:23  
Umm.

 **Tuhin** 1:35:23  
You guys had a like, you know, in a very frank exchange.  
That's good, I think.

 **Speaker 1** 1:35:26  
Umm.

 **Tuhin** 1:35:27  
Uh, but at the same time, and I'm a Prince, I don't wanna speak for you.  
Right.  
But if Prince is saying, like, OK, go read the logs directly, right.  
If that is something that you haven't done and in in your opinion, that's a low likelihood of success.  
And again, please don't take this negative there, right?  
Because you haven't done it right.  
It's not standard practice, etcetera, right? Whatever.

 **Speaker 1** 1:35:53  
Yes. No.  
Yeah, definitely, yeah.

 **Tuhin** 1:35:58  
Right, because that is that is a that is.  
I'm the Princess question.  
OK, the S3 cost right.  
Princeton, that is, that was kind of coming from the second point of view, right?

 **Jubaed Prince** 1:36:12  
Yes, that is from the S3 cost or S3 alternative point of view, right?

 **Tuhin** 1:36:13  
I think that was what was stemming from.  
Short if S3 is the problem, yeah.

 **Jubaed Prince** 1:36:17  
So it's basically optimization idea.

 **Tuhin** 1:36:21  
If S3 is the problem, right?  
Is that a better solution?  
I'm I'm making stuff up.  
I'm you guys know I'm not technical, right?  
I don't know if a no SQL.

 **Jubaed Prince** 1:36:31  
Now it's three is not necessarily the problem to him, the idea is, do we take that route or can we skip that?

 **Tuhin** 1:36:35  
Yeah.

 **Jubaed Prince** 1:36:38  
So yesterday is not a bad solution, it's just.  
I was thinking if it is possible to skip that altogether.

 **Tuhin** 1:36:41  
Yeah.  
So maybe maybe another solution?  
And again, I'm spitballing it right?  
Maybe another solution is then we we do have to create a separate app that if the like hypothetically if Divisum Kafka that CDC layer goes down for an hour right?  
We do need to maybe watermark something or whatever and there has to be a special processing.  
Maybe that's the solution that solves that, yes, three problem, right?  
I don't know.  
I'm just talking through that.  
So what I think would be helpful.  
Priyanka, before we jump on, I know we have that discussion, but based on whatever we think even internally right, we would have to manage it.

 **Speaker 1** 1:37:23  
Umm.

 **Tuhin** 1:37:24  
So I would want Sachin and Rajesh to and Tushar to jump in and say like, OK, right, this is something we can maintain.  
This is something we can learn, right?

 **Speaker 1** 1:37:34  
Umm.

 **Tuhin** 1:37:35  
Because while the solution could, but you know, I'm not looking for 100% optimized solution, I'm looking for a solution that's not going to break and that's something we can maintain.  
Sorry to put it very bluntly.  
Right.  
And and again, I'm talking from both.

 **Speaker 1** 1:37:50  
No, I think that, yeah, that that is actually a very clear you know objective to have I think thing that makes sense.

 **Tuhin** 1:37:51  
Both teams point of view, right? Yeah.

 **Speaker 1** 1:37:59  
So if you are already doing Prince Karthik, we can quickly and then we can quickly 10 minutes discuss it.  
Or what we have like this?

 **Jubaed Prince** 1:38:06  
Yeah, let's take a break and like maybe discuss among each other, break the rooms or whatever and reconvene and just come up with a couple of solutions.

 **Tuhin** 1:38:06  
Yeah.

 **Jubaed Prince** 1:38:17  
And I think Matt is also here.  
So we'll be able to make a quick decision which approach we're taking.

 **Tuhin** 1:38:24  
Yeah.

 **Speaker 1** 1:38:24  
So do you want me to create a breakout room?

 **Jubaed Prince** 1:38:28  
Whatever works, right?

 **Speaker 1** 1:38:28  
So so are you want to continue here then?

 **Tuhin** 1:38:29  
How very good Soni there.  
Or do it like you know you.

 **Jubaed Prince** 1:38:30  
Whatever.

 **Tuhin** 1:38:33  
That's that's.

 **Jubaed Prince** 1:38:33  
Whatever works.

 **Tuhin** 1:38:34  
That's fine.  
I think I'm.  
I'm just gonna go on mute and we'll come back and UH-1250 stern.

 **Jubaed Prince** 1:38:37  
I will go to.

 **Tuhin** 1:38:41  
That sounds good.

 **Jubaed Prince** 1:38:42  
Yep.

 **Tuhin** 1:38:42  
I don't know in India that is going to be.  
I believe 1020 will come back at 10, right? Yeah.

 **Speaker 1** 1:38:47  
It's OK.

 **Sachin Naikwadi** 1:38:49  
Enter.

 **Speaker 1** 1:38:50  
OK, so I think most of the technical people are here and leave it there and we can start discussing here right against so.

 **Tuhin** 1:38:55  
OK.  
OK, perfect.

 **Speaker 1** 1:38:59  
So let me put out like we have three options like let's give two minutes to each and we can all jump and then we can time bound that.  
OK, so let's first take the right.  
So, umm, rather I'll go other way around, right?  
Let's take the debezium cafka and we have the Lambda function directly fetching it from real time and getting it processed and putting in right.  
So pros from my side, I think Pro is that obviously it's going to keep working that it's real time, it's processing, it's putting into India reporting database, you have everything ready.  
Palpated falls on my side.  
I feel that it will take a little time not as much time.  
It's OK in that that is not even a con, but it'll keep running all the time, right?  
So we'll have to look and there is certainly restrictions on Nando's, right.  
I think we will anyways have to as we go into production as Prince you said earlier, we might have to put it in containers right to deal with everything.  
So that is there anybody has any points for this approach?

 **Karthik Raghu** 1:40:09  
Yeah.  
Yeah.  
See, even with the containers, we have to make sure that that particular I should functions also support this kind of mechanism.

 **Speaker 1** 1:40:14  
Umm.  
Yes.  
Yes.

 **Karthik Raghu** 1:40:20  
OK.  
I just talk like from this, right?  
So whatever solution that we provide to you just need to make sure that we, yeah.  
And well as you think is available, I think we are only talking about AWS here, but I think please consider the on the fact that we are like you know the long term solution would be on like would be sustainable long should be sustainable.

 **Speaker 1** 1:40:36  
Yes.  
Yes, definitely.  
Parthiban I think on that point we already discussed, right that when we put it in Lambda, we should not be putting it like 2, just fail out Lambda rather put it in either Python pyspark which is direct code lift and shift code right into Azure functions.

 **Karthik Raghu** 1:41:06  
Yeah, that's.

 **Speaker 1** 1:41:09  
The only on the configuration side little.  
Yeah, I think that's not a big lift and shift or configuration on functions so that I'll put other assumption before even developing such an approach.

 **Karthik Raghu** 1:41:23  
Yeah.  
Well, whatever you do, you test it on both the platforms.

 **Speaker 1** 1:41:25  
So umm ah.

 **Karthik Raghu** 1:41:27  
That way I think we can instead of doing this hypothesis, we we are sure that this this will work on both the platforms, right?

 **Speaker 1** 1:41:31  
Umm.

 **Karthik Raghu** 1:41:34  
So that is what I would suggest.

 **Speaker 1** 1:41:37  
Umm make make sense?  
OK.  
But umm, so these are all the things that we'll, you know, keep in mind when doing any longer functional, even writing the request for code.  
Right, But what do you think about the approach?  
Because I wanted to hash out all three approaches and see which one we should start working on 1st as a first priority that we think this is a better approach on trade offs on cost, everything right so.  
Touching Anand Sathish, you guys have any input on doesn't this approach?  
What do you think about Lambda consuming directly from the Kafka topics and calculating computing and protecting?  
This is 1.

 **Anand sivaraman** 1:42:26  
Uh, so.

 **Speaker 1** 1:42:26  
Any pros and cons?

 **Anand sivaraman** 1:42:26  
So Priyanka, I think I think I think going back to both karthick's and tune Karthik and Tuhin and Princess Point, I think we need to we need to identify the top two or three items to experiment and very quickly experiment it by today or tomorrow because every every solution has its pros and cons, right?

 **Speaker 1** 1:42:38  
Umm.

 **Anand sivaraman** 1:42:45  
We need to work.  
What is clearly applicable here?

 **Speaker 1** 1:42:46  
Umm.

 **Anand sivaraman** 1:42:48  
So obviously I see true priorities.  
What are the options other than S3 that we have to take if we are, you know, how do we, how do we bring in that fault tolerance without using S3?

 **Speaker 1** 1:42:54  
Umm.

 **Anand sivaraman** 1:42:59  
The second is obviously within the with the within the Debezium Kafka and Lambda approach, how can we handle this right?  
I think these two can be very quickly experimented within a day or two and we can move forward accordingly.

 **Speaker 1** 1:43:09  
Umm.

 **Anand sivaraman** 1:43:13  
Bring up because time is of essence here and obviously we cannot keep on experimenting.  
So let's pick up something that is doable in a day or so, and then report it accordingly.  
I can't, OK.

 **Speaker 1** 1:43:32  
I think our own well, I agree to that.  
So I think Prince is not here, but I wanted to bring up that point that that is the same thing that I'm concerned about.  
It's not that sin log as doing also mentioned right that bin log we cannot read or we that is time consuming and even on the other side, what do you think Anand that this is something that will have to be looked into continuously manually we can definitely test it out in the third approach right.

 **Anand sivaraman** 1:43:58  
So.  
So we see in the in the in the in the past at least wherever I have seen architectures, Debezium, Kafka and Lambda have always worked right, like the only thing is are we the?

 **Speaker 1** 1:43:59  
We can test out.  
Umm.

 **Anand sivaraman** 1:44:12  
Are we ruling out the fault tolerance that S3 can bring in?  
If that means, then then the, then the that the pipeline that we are building without is 3 has to be so robust and all the retry mechanisms have to be good so that we don't lose data.

 **Speaker 1** 1:44:20  
Umm.

 **Anand sivaraman** 1:44:24  
So that is that is the only thing and I think for for this purpose.  
That's why I kept repeating option A and option.  
See, that's to ensure both work.

 **Speaker 1** 1:44:33  
Umm.

 **Anand sivaraman** 1:44:35  
Only way for us to really confirm that is we can do a very quick experiment.  
Right.  
And and Kafka has been a proven technology.  
So it's not like Princess saying the the there is no issue about Kafka.

 **Speaker 1** 1:44:44  
Yes.

 **Anand sivaraman** 1:44:47  
The thing is, can we?  
Can we do without S3 and that is possible?  
That's not a problem.  
Reading through Bin Log is also possible if it if it can read from any file it can read from bin log, right?  
That's not an issue also.  
Yeah.  
So we can we can take it up.  
Uh, Priyanka.  
And then we can very quickly explore it, get this sorted out, yeah.

 **Speaker 1** 1:45:17  
Yeah, I think it will be better that we look and all three approaches we compare it rather than, you know, talking about it.  
What do you think?  
I think that makes sense and hopefully by the end of the day.  
Think about what is possible.

 **Anand sivaraman** 1:45:31  
Yeah, I I I.

 **Speaker 1** 1:45:32  
How much it will take time.

 **Anand sivaraman** 1:45:39  
Or at least the first two we should be able to try it out.  
See, it's not about the feasibility, right?

 **Speaker 1** 1:45:42  
Then.

 **Anand sivaraman** 1:45:43  
It's about what what works for us better for this particular situation with all the, with all the vision that we have so far.  
See, I think that's probably what we need to confirm all of these solutions will work.  
What works for us better is what we need to be very clearly calling out.

 **Speaker 1** 1:45:59  
Umm.  
So we'll Karthick one question I have for you because I know that the Prince has a concern about S3 in Series C and Anand.  
Please correct me if I'm wrong.  
In CDC, we will actually be having one copy of the entire, you know data.  
And so whatever your data.  
Ohh past right now you I I understand the concern having that same data duplicated there, right?  
Hey, is it as much cost right now or when you, uh, take a backup of the AWS RDS, right?  
That is the same kind of cost you'll have for the S3II, although I should correct myself in the in the backup you take a snapshot, so it's compressed, but in this.

 **Karthik Raghu** 1:46:49  
No, I think yesterday is stress cost is negligible.  
I don't think like we have ways to handle the data.

 **Speaker 1** 1:46:53  
I E.

 **Karthik Raghu** 1:46:55  
It's request is very negligible to what we are paying today. Uh.

 **Speaker 1** 1:47:02  
Actually that was the same thought I have.

 **Karthik Raghu** 1:47:07  
It's very maybe.

 **Speaker 1** 1:47:07  
OK.

 **Karthik Raghu** 1:47:08  
I don't think it's mostly, I think TV's of data costs $5, so.  
We became just talk about live site Team management as well to in order to what I like to handle the life cycle like.  
You know what we wanna do, right?  
You know, waiting on the access point is like something like glacier.  
And there's also parking, right?  
There's been probably serious anymore.  
Thank you.

 **Speaker 1** 1:47:38  
The same.

 **Karthik Raghu** 1:47:38  
You know, just just a few cents.

 **Speaker 1** 1:47:38  
Yeah, exactly the same that we discussed earlier, right here, we are going to have the lifecycle management, making sure that it does not have unnecessary copies, unnecessary data and it is just a stop gap between, right for a fail safe.  
And honestly, if you ask my opinion and Prince, please feel free to comment on that.

 **Karthik Raghu** 1:47:52  
Yeah, yeah.

 **Speaker 1** 1:47:59  
We log uh option that you're saying is not like reading directly is not like it's not feasible.  
Let us feasible you can do it, but then you have to deal with all the developments of putting out the code that whatever you're reading and you're dealing with, it should be correct.  
And reading through the code and putting you know all the trucks like making sure in the CDC you go, I look to about it and binary logs you can go read it like you can have an API fetching through it but that is a development time.  
Maybe you have to make sure that we have everything correctly implemented, right?  
The general practice is to have debezium Kafka or you know AWS.  
Whichever kinesis is a little on the cost side, I wouldn't say for that the Museum Kafka running and now if we talk about whether I have to have S3 in the middle or not, we can look, we know that S3 can be the you know the thing that is not a problem in implementation.  
S3 tool Lambda also what you were saying earlier, right?  
We can fetch it every minute so that we don't run Lambda unnecessarily to cost because it's serverless.  
The other option is the having debezium coca setup with the with your Lambda function.  
My buyers, as we call right, let's call it Baskar functions, right?  
Application and directly consuming from Kafka topics transforming and putting in into the in your reporting DB.  
So that is the other option which we can test out and we can see if it's it has any you know any restrictions.  
I would prefer these two approaches.  
It's being you can read it simple API, but then that development time and then maintenance like that is something I'm not very keen on honestly because either it's my team or your team doing it later.

 **Jubaed Prince** 1:49:54  
So.

 **Speaker 1** 1:49:57  
And then and development time is what is, you know, stopping me from suggesting that although you that will be a new product you have because that is why people use Division Kafka right now because it deals with all of them.  
All you have to do right now.  
Like what I was talking about crowds, right?  
All you have to do is make sure that the delete which Kafka actually deals with right in the.  
In CDC, you just have to make sure that you are also considering that using Kafka here you'll have to make sure that for all the logs that you're reading, when events are that are happening, you deal with.  
OK, this was deleted hardware and get that deleted data into.  
Why this thing into my new DB or delete it from there?  
Kind of.  
So you have to consider all the scenarios.  
So what do you think?

 **Jubaed Prince** 1:50:44  
Yeah, I get that.  
So what was the this final decision on 2/2 approaches that we are you are gonna experiment with and does Annex Cloud team has buying into that?

 **Speaker 1** 1:50:57  
So actually I think it's just me and Karthik talking about it.  
But what do you think Karthik?

 **Jubaed Prince** 1:51:02  
OK, so Karthik is Annex Cloud, right?  
So tell tell us that two approaches that you decided.

 **Speaker 1** 1:51:05  
Yeah, that's why I'm putting you.  
Ohh Karthick are you OK with the earlier ones?

 **Jubaed Prince** 1:51:11  
And again, I'm not married to any of these solutions, right?  
These are all to put on the table to discuss, right?

 **Speaker 1** 1:51:15  
Sure.

 **Jubaed Prince** 1:51:17  
We can always go back and figure others out if this fails, so it's good that we we have to pick two.

 **Speaker 1** 1:51:18  
Yes.  
Yes. No.

 **Jubaed Prince** 1:51:24  
So let's pick two and go ahead with it.  
Since we don't have time, this is the benefit of having no time.

 **Speaker 1** 1:51:29  
So.  
Yes.  
So I would suggest the Karthik please the tell me if you agree or disagree.  
I think first will be we should actually have for now S3 cost is not as much right to have one copy of the data.  
It's CDC.  
It's not going to write the whole data, so it's not.  
It's very negligible compared to what you are.  
You know, thinking right now is there so and this will give us a fail safe.  
Also, there'll always be a current copy and the reads that are there, right?  
Because it's updates from Lambda function, it'll trigger, right?  
Yeah, these are the new updates.  
Let's deal with it into the leaving, OK, so the first I would say the first priority would be to test that out.  
If we think that, OK, this is costing or this is something we can on the second approach that we should go with is again the Kafka debezium and directly fetching it using the patch back port right on Lambda and then we can deal with it in containers or you know and putting it into reporting DB.  
What do you think, Prince?  
Karthik, do you think this is good priority?

 **Jubaed Prince** 1:52:38  
Uh with Annex Cloud has no complaint to it.  
The Dev developers in engineers were present here.  
Then you should be able to move forward with this when I think this is the time to come up and say whatever you have in mind about these two approaches, or any third approach that you want to defend against these two.

 **Speaker 1** 1:52:49  
Umm.  
Umm.

 **Tuhin** 1:53:00  
OK, I hope you can see my screen.  
Can you just make sure that I'm writing it correct?

 **Speaker 1** 1:53:03  
Yes.  
Yes then.

 **Tuhin** 1:53:06  
So this is the approach one is if I if I write it correctly.  
So this is going to still going to be Kafka and Visium, right?

 **Speaker 1** 1:53:14  
OK. Yes.

 **Tuhin** 1:53:15  
So these two are still gonna be there.  
Lambda still gonna be there, but there's going to be an S3.

 **Speaker 1** 1:53:19  
And.  
Yes.

 **Tuhin** 1:53:23  
I'm just gonna put it as a.

 **Speaker 1** 1:53:23  
That's OK, yes.

 **Jubaed Prince** 1:53:24  
One has three, one has noise.  
Three, that's that's that 212.

 **Speaker 1** 1:53:27  
And and yeah, and the other thing that in the first approach I'm saying is that the Lambda function runs on the schedule of one minute rather than, you know, continuously running in the approach be it'll be, you know, the Kafka topic will actually initiate trigger the Lambda function directly and it'll contain it like how your time it'll be running.

 **Tuhin** 1:53:28  
OK.

 **Jubaed Prince** 1:53:36  
Right.  
So one thing to call out.  
Sorry Priyanka here is I think Tuhin.

 **Speaker 1** 1:53:51  
Umm.

 **Jubaed Prince** 1:53:51  
You should understand this very carefully.  
Is the reports will be one minute delayed, so delayed as in the data.

 **Tuhin** 1:53:57  
Yes, yes, yes, I got it.

 **Speaker 1** 1:54:09  
And then right, we can do right on an update in the S3 and the Lambda runs.  
But then ohh, there's three will only be a fail safe.  
This actually make sure what Prince you were calling, right?  
One minute we call the Lambda Anand Transcend 1 go right so we are not just a on every single you know one data point update.  
We are running the Lambda.  
That's what I'm thinking.  
So are you good with that Prince Tuhin?  
Like we schedule it on one minute, the land order you want to have it like the file lands a new update lands into S3 and Lambda.  
There's two type of Tejas here.

 **Jubaed Prince** 1:54:46  
I I vote on approach a I vote on approach a that seems more stable to me and cost friendly to me, so that's that's my boat is on a.

 **Speaker 1** 1:54:46  
I would prefer this one.  
That's why I suggested.  
Yes.  
I also go by.  
I also said this time was my preference would be anybody else.

 **Tuhin** 1:55:10  
I don't know enough, but I can tell you I can live with a delay of one minute.  
So if that's what you're looking for, I can live with that.

 **Jubaed Prince** 1:55:18  
Thank you.

 **Tuhin** 1:55:18  
So that's that's that's that's my part.

 **Jubaed Prince** 1:55:18  
Them.  
That's generous of you.

 **Speaker 1** 1:55:22  
Yeah.

 **Tuhin** 1:55:23  
So you guys can figure out the technical bid?

 **Karthik Raghu** 1:55:25  
So I have a question.  
So like option A I think is what we had discussed earlier, right?

 **Tuhin** 1:55:26  
Yeah, yeah.

 **Karthik Raghu** 1:55:31  
So right now we are saying like let's do option A and then option B and then CI explore.  
But then I think like majority of the team and I bet lean on to option A.  
My question is like what happens if if I reboot after server or like you know if it is down for about 1520 minutes?  
Ohh then and what happens?

 **Anand sivaraman** 1:56:02  
So, so so Karthik that that's that's the reason why you know I also mentioned 2 for the Kafka servers not to be down.  
Umm, you know if we can go through AWS MSK manage services, Kafka, AWS manages all the scaling and you know the troubleshooting just like any other AWS service country.

 **Karthik Raghu** 1:56:23  
Yeah.  
Again, you're talking on on on again.  
You're talking on.  
Is something specific to AWS.

 **Anand sivaraman** 1:56:31  
No, this is this is just the man.

 **Karthik Raghu** 1:56:31  
Right.  
So do we have?  
Do we even know we have?  
Yeah.  
Yeah, that's fine.  
But do we even know?  
Like, do we have a similar service on Azure?

 **Anand sivaraman** 1:56:37  
Umm.

 **Karthik Raghu** 1:56:39  
Because like whatever we provide on AWS, we should be able to provide the same on Azure right? So.

 **Anand sivaraman** 1:56:41  
Yeah. Yeah, yeah.  
Yes, yes, Sir. Yes.  
I as Azure has Azure, Azure has managed services as well Karthik.  
That is, I just told this as a as an alternative, but we can ourselves maintain a craft a cluster the just like we are maintaining any server we can maintain, but that that needs additional like like you mentioned putting in Splunk or whatever we can typically do what we do with any other server but the the chances of.

 **Karthik Raghu** 1:56:52  
OK.

 **Anand sivaraman** 1:57:14  
You know, leaving that to AWS or Azure is is a little more easier for us than doing it ourselves.

 **Karthik Raghu** 1:57:20  
I've Greene agree.

 **Anand sivaraman** 1:57:20  
That's all I'm trying to say.

 **Karthik Raghu** 1:57:22  
If we really wanna go with Manish SERVIVCES, then obviously we don't need S3, right?  
There's no point in because they they also like if it's managed services they provide 99.9 or maybe 9-9 percent up time and then they will also say like give some metrics for the data reliability as well.

 **Anand sivaraman** 1:57:29  
In in, in.  
Correct.  
Correct Karthik.  
Correct, correct, correct, right.

 **Karthik Raghu** 1:57:42  
So in that case I don't need S3, so I can just go directly query directly with the Kafka.

 **Anand sivaraman** 1:57:44  
That you're you're.

 **Karthik Raghu** 1:57:47  
Let's go with option B, then, right?  
So that's my point.

 **Anand sivaraman** 1:57:50  
You're you're.  
You're you're absolutely right.  
The only reason why I brought in that S3 thought process Karthik is it's a it's a convention that is typically being used when when we use streaming data that is #1 and and where we have seen with other clients.

 **Karthik Raghu** 1:58:00  
OK, right.

 **Anand sivaraman** 1:58:04  
Karthik #1 #2 is there's three cost.

 **Karthik Raghu** 1:58:06  
OK.

 **Anand sivaraman** 1:58:07  
Like you said, it's negligible.  
Harish Year off security, that's all I was looking at.  
Nothing, nothing more, nothing lesser.

 **Karthik Raghu** 1:58:18  
OK.

 **Jubaed Prince** 1:58:18  
And also I want to add here that I since POC stage I would feel comfortable in a managed service over we managing the Gupta servers and also thinking in terms of Annex Cloud that we have no almost no experience in managing Chapter Kappa servers.

 **Karthik Raghu** 1:58:29  
Yeah, yeah, absolutely, absolutely.  
But yeah.

 **Jubaed Prince** 1:58:37  
So maybe a managed service is the way to go here.

 **Karthik Raghu** 1:58:40  
Yeah, absolutely.

 **Anand sivaraman** 1:58:41  
Yeah, I'm yeah.

 **Karthik Raghu** 1:58:42  
Totally agree to it like, yeah.

 **Anand sivaraman** 1:58:43  
And I'm yeah.  
And and adding to that Karthik, the AWS kind of sometimes confuses as well.  
I providing two different solutions for the same problem right?  
Like for example they have managed services Kafka supporting Apache Kafka and they also have their own tool.  
So if we go there Kinesis route, we are very much tied into AWS, but at least if we go the managed services cuff, managed services, Kafka route, we have an opportunity to use that in Azure also because Azure also provides that capability.

 **Karthik Raghu** 1:59:03  
Maybe that?  
Then.  
OK.  
So then do we have any idea about the costing?  
Because I'm absolutely not sure.  
Like you know what?

 **Anand sivaraman** 1:59:21  
We we'll we'll create that report.

 **Karthik Raghu** 1:59:22  
What the cost will do is.

 **Anand sivaraman** 1:59:23  
We'll create that we will create that report and get back to you, Karthik, both on AWS and Azure will get back to you.

 **Karthik Raghu** 1:59:29  
Focus.

 **Sachin Naikwadi** 1:59:31  
Uh, Anand.

 **Anand sivaraman** 1:59:32  
But Priyanka.

 **Sachin Naikwadi** 1:59:33  
One question here, what about DBZ and failure?

 **Anand sivaraman** 1:59:41  
That that can that can happen, Rajesh, like I I think that's something that we need to monitor the and that there needs to be some level of monitoring and processes that we have to put in.  
So that is something that we'll explore as part of that and see what can be done there.  
Yeah, I.

 **Tuhin** 2:00:05  
OK. Umm.

 **Anand sivaraman** 2:00:05  
We have.  
We have seen we have seen scenarios where debezium also has failed.  
But see in a pipeline that can be any any step that can fail the point is not about the failure.  
The point is about how effectively we are able to retry that same transaction, right?  
So that is something that both Kafka and Debezium provides because of the Kafka topic capability as well as debezium.  
So that is the hope with which we bring in these topics, but we will explore, add more pointers to add as so well in all these areas, right.  
So Priyanka Rupali, let's take a note of this.  
Add it to the Monday board and make sure that we'll answer these questions properly as well.  
So great questions.  
Thank you so much for bringing this up.

 **Speaker 1** 2:00:44  
Sure.

 **Tuhin** 2:00:50  
OK.  
Did we just talk ourselves into only one approach, which is great by the way?  
So just wanted to make sure. OK.

 **Jubaed Prince** 2:00:56  
Yes, congratulations.

 **Speaker 1** 2:00:58  
Yeah, yeah.

 **Jubaed Prince** 2:00:58  
We have come to an agreement.

 **Tuhin** 2:00:59  
OK.  
OK, good.  
OK, good, good. OK.

 **Jubaed Prince** 2:01:02  
And we are two minutes late, which is also good.

 **Tuhin** 2:01:06  
OK, OK, sounds good.  
If I know I'm I I want to be respectable for everyone's time, but give me just one second.  
I I wanna recap the board before we break, OK, so.

 **Jubaed Prince** 2:01:21  
And also to hit can you go through tomorrow's workshop agenda so that like everyone has some idea about tomorrow's activities if if anyone wants to put something there?

 **Tuhin** 2:01:29  
Sure, sure, sure, sure, sure.  
Yeah.  
Yeah.  
So data model, we talked about it, this is going to continuously get improved worksheet technical worksheet.  
We're working on it.  
We're close to being done.  
Initial data migration.  
Is this something uh we have to do soon, Sachin from.  
Uh for the points you port based on the report, that data model that you have created, is that something you're planning on doing or is that something we will keep on hold similar to who are due date for the?

 **Jubaed Prince** 2:02:06  
I believe Sachin is on another call to him, but I think I think has already done this.

 **Tuhin** 2:02:09  
OK.  
For interaction report, not for the points report.

 **Jubaed Prince** 2:02:16  
All right.  
Right for point strip or. Yeah, correct.  
You're right.

 **Tuhin** 2:02:19  
OK.

 **Jubaed Prince** 2:02:19  
But he is in another call.

 **Tuhin** 2:02:19  
I'll.  
I'll come back to Sachin on this, OK.

 **Jubaed Prince** 2:02:20  
He just texted us, yeah.

 **Tuhin** 2:02:23  
OK, sounds good.

 **Speaker 1** 2:02:24  
So the so he know Prince overhead, please.

 **Tuhin** 2:02:25  
Establish.  
Yeah, go ahead.

 **Speaker 1** 2:02:28  
No, not the question of initial data migration.  
Do we want to do it before we like implemented anything or we want to test it out once we have the implementation of the points report and you know, run it through so that because.

 **Tuhin** 2:02:40  
Up to you up to you either is fine, right?

 **Jubaed Prince** 2:02:42  
Bumper, you know.

 **Tuhin** 2:02:43  
Because we can just delete the tables records right?  
Then just from the process again, in my opinion, in layman's opinion, it doesn't matter, right?

 **Speaker 1** 2:02:47  
Hey.  
Yeah.

 **Tuhin** 2:02:52  
But it's up to you guys.

 **Speaker 1** 2:02:52  
Yeah, we can just truncate the table.

 **Tuhin** 2:02:53  
Yeah, yeah, yeah, yeah.

 **Speaker 1** 2:02:54  
Yeah.  
OK.  
But but the I was thinking, OK, this is something we'll let Sachin decide, right?  
Anyways, it's going to work.

 **Tuhin** 2:03:00  
Yeah.

 **Jubaed Prince** 2:03:02  
I'm the question with this Lambda function.

 **Tuhin** 2:03:03  
Perfect.

 **Jubaed Prince** 2:03:05  
Language are you using to write the code inside language function?

 **Speaker 1** 2:03:12  
We'll go by Python or Pyspark, mostly spark because the spark streaming with Kafka is the, you know, compatability and you know Lily, Lily going to, you know, put it.  
No one thing that we have to keep in mind is we don't use AWS portal so that it does not lock the code in.

 **Tuhin** 2:03:27  
OK.

 **Jubaed Prince** 2:03:34  
Talk.

 **Tuhin** 2:03:35  
OK.  
Ah, anyone from Annex Cloud will do.  
We still have someone from Dev team.  
Yeah.  
No, just making sure is that a Tushar.

 **Tushar Balwadkar** 2:03:45  
Uh, you have to.

 **Jubaed Prince** 2:03:45  
Ohm those are is here, Tushar, I believe is working on a node.

 **Tuhin** 2:03:50  
Sorry to OK to share.

 **Jubaed Prince** 2:03:50  
Chairs lab as well.

 **Tuhin** 2:03:53  
Sorry to make you the A spokesperson for the Dev team, but is that manageable?  
Do we have in House expertise or willingness to learn either of this technology?

 **Tushar Balwadkar** 2:04:04  
I need to check actually by something like need to check.

 **Tuhin** 2:04:09  
OK, OK.

 **Tushar Balwadkar** 2:04:09  
I will check with the team. OK?

 **Tuhin** 2:04:11  
Can you can you can you take that as an action item, please?

 **Tushar Balwadkar** 2:04:11  
Anyone else? Yeah.

 **Tuhin** 2:04:13  
Yeah.  
Yep.

 **Tushar Balwadkar** 2:04:15  
Uh, yeah, yeah.

 **Tuhin** 2:04:15  
OK, perfect.

 **Tushar Balwadkar** 2:04:15  
I will update you. Thank.

 **Tuhin** 2:04:17  
Perfect A2 shirts trying out and not just application that he can separately that that is OK that can continue.

 **Jubaed Prince** 2:04:24  
This will also be hosted in a Lambda function, but it's it's based on JavaScript.

 **Tuhin** 2:04:30  
Yep.  
OK, fine.

 **Jubaed Prince** 2:04:31  
It it's gonna do the same thing that the previous one is doing.

 **Tuhin** 2:04:35  
OK, then I a important question.  
Can he need him?  
Do you have access to whatever you need to have access to to run the experiments?

 **Speaker 1** 2:04:49  
Then you brought up.  
I was going to bring up the same thing.  
We don't have access to.  
We don't, yes.

 **Tuhin** 2:04:53  
OK.  
So Kartik, I will create this for you and Karthik and Priyanka, if you guys could forget that out today, that would be great.  
Karthik, maybe you can the recommend someone or.  
Karthick here still or no?

 **Karthik Raghu** 2:05:11  
I think we'd have to give them access so to the portal we would do that, yeah.

 **Tuhin** 2:05:12  
OK, OK.  
OK, got it.  
I will put that as an action item for all the Priyanka got to give you.  
That's good.  
Just coordinate today.  
Yeah, don't become a blocker, OK?  
Or tomorrow.  
Then I'm hoping we don't have to come back to it.  
We can, uh, where we know the experiment that we're going to run, we can talk more about the data model.  
If there is questions, uh then the centrally let me just go back to me one second.  
I'm gonna stop sharing and I'll come back to it.  
So visually, maybe this will give us kill so.  
I feel like we have figured out this part today to an extent.  
Some tables within here, right?  
So if I look at the remaining work in terms of discovery, in my opinion and here I think Prince, you may have to help out as well.  
And anyone please feel free to chime in.  
I think there is a concept of a materialized view.  
We know my SQL doesn't support it natively, so we may have to debate about pros and cons, and there is a separate I think that would have to talk about, which is essentially for the exports optimization and if there are tools, technologies etcetera, we can use, right?  
What kind of process we're gonna flow and that is pretty much it on a high level kind of fraud, the end to end.  
Ohh and to end approach there are some uh patterns of Castro donation and test plan creation.  
Ah, other than that, I think we might be good for tomorrow's session.  
I don't know if anyone else wants to add any agenda item for tomorrow.

 **Jubaed Prince** 2:07:22  
Yeah.  
Just want us to maybe think through everything that we talked about today and maybe spend some initial time on discussing yesterday and in between if there was any work.  
And then we we get into all this topic, but I think heavily it is we should discuss and think about test testing side of as the edge cases corner cases fail saves and all those things and special also security.  
So let's let's discuss all those items tomorrow probably.

 **Tuhin** 2:08:03  
OK.  
Priyanka sounds good.

 **Speaker 1** 2:08:07  
Yeah.  
Well, yeah, that sounds good.

 **Tuhin** 2:08:09  
Yeah.

 **Speaker 1** 2:08:09  
I have one question.  
I know I'm jumping to it and materialized views.  
Prince, do we need that if we are having a reporting DB that will directly like our reports are calculated in the way we don't, do we still need the materialized views on that?

 **Tuhin** 2:08:13  
Yeah.

 **Jubaed Prince** 2:08:21  
For.  
Umm well it it will be fast in my opinion without it.  
But if you put that in, it will be even faster.

 **Speaker 1** 2:08:31  
Umm.

 **Jubaed Prince** 2:08:32  
So it's it's really the question how how far you you wanna take that?  
So it's it's really up to you guys, but this I would say is the ideal solution in our opinion.  
But again, we have time limitations and everything.  
So so I I think we can discuss that tomorrow.

 **Speaker 1** 2:08:49  
Umm. Chuck.

 **Jubaed Prince** 2:08:50  
These are all discussion for tomorrow either.

 **Tuhin** 2:08:54  
Yeah.

 **Jubaed Prince** 2:08:54  
All, all the fight again tomorrow.

 **Tuhin** 2:08:55  
And there's some use cases for some external.  
Try to explain further tomorrow as well.  
Maybe their own spark?  
Some ideas?

 **Speaker 1** 2:09:12  
Sure.

 **Tuhin** 2:09:13  
OK.

 **Speaker 1** 2:09:13  
I think that works.

 **Tuhin** 2:09:13  
OK.  
Yeah, sounds good.  
Anything else from anyone before we drop?  
Affect.  
Thank you so much everyone.  
Thank you for your time and thank you for the discussion.  
I appreciate everyone being very open and sharing their thoughts freely.  
Thank you. Yeah.

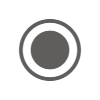
 **Speaker 1** 2:09:39  
Thank you everyone.

 **Muniaraj Thangavelu** 2:09:39  
And zero.

 **Pranav Marathe** 2:09:42  
Thanks.

 **Jubaed Prince** 2:09:42  
Thank you all, bye.

 **Tushar Balwadkar** 2:09:43  
Angel.

 **Muniaraj Thangavelu** stopped transcription