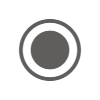
**Annex cloud-Kanini\_ joint Working session-20240322\_083217-Meeting Recording**

March 22, 2024, 3:02AM

2h 15m 1s

 **Priyanka Kochhar** started transcription

 **Tuhin** 1:04  
Just been couple of people.  
Just give me a minute or two.  
If not, we'll get started.

 **Speaker 1** 1:11  
Shut it.

 **Tuhin** 2:15  
OK, Penka Prince, let's get started.  
The Tim just will have to be caught on and you know we can share the recording and we can we can bring them up to speed pianka just and and and canine team just so you're aware we tried breaking down the project as much as we could and thinking what are some of the steps we might have to do.  
So Princeton, I came up with some tasks.  
We're not saying like these are the has that we have to do and we started thinking about like you know how we can allocate.  
The respective team members from each other's team and you know these are just proposal at this point.  
So the goal is to go through, see if the plan is sound and then add or mid item.  
Does that sound good as a starting point Venkata?

 **Anand sivaraman** 3:15  
Sure.  
Uh, dude, sure.  
Yeah, we could.  
We could start with that.

 **Tuhin** 3:20  
It's gonna go ahead.

 **Jubaed Prince** 3:23  
Yeah, just give me a minute.  
Am I visible at my screen visible?

 **Speaker 1** 4:23  
Yes.

 **Jubaed Prince** 4:24  
Cool.  
Umm, so we essentially divided it into two parts.  
One is for this end of the week I experiment, plan and then for umm next week we're trying to look into a staging plan which is kind of like a productionalized version of the experiment that we are doing this week.  
Umm so it's starts with our team already worked on debezium setup in the US E One database which is basically the AWS database that we are using as a primary source of data for us.  
They created a MySQL 8 RDS database which is the report database. Uh.  
And.  
And then a report, which is the interaction report has been already created in there, which is kind of like a aggregated table as part of like a POC and they have all also migrated data from US E 1 to the MySQL 8 RDS.  
Next part, which is still we have to work on is having another application which receives the CDC events and they are going to copy the data from US E 1 to the MySQL 8 RDS.  
Umm, next what we do is we are going to kind of do the same next week and probably for the thinking data sync mechanism, we'll have to figure out the option that you suggested the Lambda function or is there option to explore node JS application or any PHP application given that PHP is like the primary language that our team is used to.  
So I think we can have certain discussion around that.  
Umm, we also would potentially need a data generation script which could be used to generate random data for the points report.  
By random I mean the data will be random, but it will be structured like real data so that we can check the CDC part and the real time upgrading parts.  
Umm.  
And we also want to make sure before we jump in and do that staging plan, we are going to who are points report analysis.  
What I mean by that is we will look into the database structure and schema and we will try to answer some of these questions like are those data?  
Are there any data points that are inconsistent?  
Are there points that's causes delays, redundancy?  
Are there conflicting situations for any data point in future?  
In future, if we have to make some schema alteration, how difficult is that in terms of creating a new column?  
Or rename a column.  
There can be other scenarios as well.  
Umm, also we we want to know, will there be any challenge if there is a bug in generating bug we find in a computed column.  
So in future, what are the processes involved in order to tackle those?  
The goal here is to really question everything about the data in the reports in the points report and kind of come up with improvement suggestion and also in terms of schema index in stored procedures, metal as views and views.  
So my SQL native stuff, what can we do to optimize things?  
We also need to ask do we need to rethink how application layer provides are being done for the data?  
Do we need any sort of change there?  
And also we want to kind of forget the current implementation and think from a very first principle whether there is if we have to freshly do some of this work, is there any groundbreaking thing we can do in order to drastically improve performance of this particular report?  
Umm.  
And then, uh, these are the people that we are interested and that these people will do this analysis, maybe it can be a synchronous work like scheduling 30 minutes and going through it or it can be asynchronous where we all pitch in our input in a worksheet and also we want to analyze angular application optimization opportunities as well.  
So in the next week, we also want to set some stretch goals if we are able to do one to six, we want to see if we can work with the export excel function and creating automation test scripts to test the results that we see.  
Uh, so those will be something we want to we would love to explore next week as well in terms of bigger milestones, we next week we are going to by the end of next week, we're going to have our opportunity to measure the performance of points report then then we will also try to align our leadership with the new approach that we find.  
Uh, we'll have a diff between existing report and new report.  
Maybe it will not be as fast or it will be better than before.  
We don't know.  
So those will be the things that will be looking into by the end of the next week.  
Then we'll want to create recipes to rebuild the 70 reports.  
From what we learned so far and then the next week after March 29th would be focusing on forming a team that can follow that recipe and do 17 reports kind of parallelly and hopefully by the end of April we have a rollout plan for all remaining reports.  
Umm and.  
In terms of task, we want to turn off data bricks from US AP Region and also on investigate the minimum tile and requirement for Harrods use case so.  
Here I tried to form a data flow diagram which is let's said Harold's it's loyalty data are going to as usual go to US one uh.  
But from there a change capture event will go to some sort of node app in Lambda.  
Let's just say that's the process we go with and then it will generate some aggregated report and say within the report DB.  
On the other hand, have received the search for a points report with them and they trench.  
Then it goes directly to the report DB, gets the Harrow data within the date range and shows in the area.  
So from a high level, this is essentially what we're doing on next week.  
Tuhin, juana.  
Add something here.

 **Tuhin** 11:51  
Everything is accurate apart from the herd space.  
Let's maybe maybe we'll come back to it, but.  
I think on a high level that's kind of the thought process right now.  
So just want to hear from the team he if we go up the steps like does this makes sense?  
Are we disagreeing with anything at their steps that are important?  
So maybe it's better to open up coming back now from top down and let's see how much progress you can make.  
Obviously like this is just a proposal at this point, so it just really need to understand if you have, if the team has additional viewpoints are things that we haven't considered or things that may be unnecessary.  
So maybe we start there.

 **Anand sivaraman** 12:46  
Yeah, but Tuhin/Prince, thank.  
Thank you so much for putting through this.  
I I think well thought out.  
Thanks again.  
Just a couple of points, maybe I might have missed.  
When Prince was going through this one was a the the assumption that we are going with is there are no changes that are happening to the existing MySQL queries.  
Umm.  
In in terms of any business logic change or any performance tuning change that is happening in parallel while this exercise is going on, that's that's an assumption I would assume.  
And the second thing is for the testing of this exercise in terms of CDC debezium, I would also assume that we will, we will need to bring in production like data into lower environment for us to test a.  
So just wanted to understand if that can be added so that like we are clear about the, the, the steps involved.

 **Tuhin** 13:40  
Yeah.

 **Karthik Raghu** 13:42  
So let me explain what is done Anand.  
So basically we took a clone database, cleaner clone of a production database.

 **Anand sivaraman** 13:48  
OK, OK.

 **Jubaed Prince** 13:49  
Yeah.

 **Karthik Raghu** 13:49  
OK.  
That is step one right?  
So we created a clone out of a production database.

 **Anand sivaraman** 13:51  
Sure, pepper.

 **Karthik Raghu** 13:54  
We created one more cluster and then in the same VPC we have created one and and machine in situation and then installed debezium plus factor in it.

 **Anand sivaraman** 14:06  
OK.

 **Karthik Raghu** 14:07  
OK.  
And then we went to the configuration like, you know, in the Debezium Connect properties.  
Finally, there, right.  
So we just modified and then just connected it to this clone database.  
For now, we have connected it into the Clone database.  
We have not connected it to product still because this is a proof of concept.  
We just have to make sure that everything is working as expected.  
It's a week benefited this Kafka instance too.  
Lonely, that is.  
And then we started like getting the streaming data from the from the.  
What you say?  
Ohh the clone database like whenever we update it, they will alter the table.  
We started getting all those events into the system, so now that what we need to figure out is like, you know, the next step is like we need to figure out.  
Ohh, you know whether we want to implement an application which will actually pull the data from the data and then push it back into a new database like this has nothing to do with the production database, right?  
This is a brand new my sequel late cluster.  
It will have a reporting data like.  
Basically it will have remember like when we were discussing the other day like we we thought that we we discussed about bringing in transformations into the data and then I ringing all the data in one place and then using it for the reports right.  
So the same thing we are doing, the collected data would actually be arranged or transformed in such a way that it is easier and faster for the reports to be generated from the MySQL database itself, right?  
So that's the idea?  
Umm, so I don't know whether you want to use you know the sink connector for the Kafka and then do it because I don't know how the data can be transformed.  
Uh, in such a way that we in into a schema that that we think would actually help, like basically the reporting schema, right, like whatever the fields are required for the reports.  
So the new application that was talking about is basically whether you want to create that application.  
Uh, are you?  
You start writing that application or you use any kind of native tools provided by Kafka.  
Something like a sync connector which you can connect to my SQL another my SQL database and then start pushing the data from prod to my SQL but then one thing has to be remembered is like we have to transform the data when we are sending.  
Ohh like the data from the CDC information or the data from the old server to the new server.  
I think we have to make sure that the data is transformed and then be ready for some kind of operations like it should be performance focused.  
So we are not doing any kind of.  
We may not be doing any kind of changes to the existing database like you know any kind of performance in.  
Gonna use that database anyway, right?  
So this this Kunta acts as a middleman to get the data, fits the data, it's the entire data and then just pick what are the pieces of information that is required and data fields that are required for the reports.  
And then pick those fields and then push it to a new database.  
So that is the plan actually.  
So you can just let me know if you have any questions here.

 **Anand sivaraman** 17:50  
I'm uh, I'm able to understand the Karthik, Ashok Priyanka.  
Any questions from your end?  
Ashok, I know that we have implemented the BBC and Kaushika and other installation.  
So anything that you want to call out?

 **Ashok KJ** 18:01  
You know, I I just could quickly want to understand the the thanks Karthik.  
I think that that was my I was about to ask that question.  
I think you answered most of them.  
So in terms of this division configuration, it's directly on the on the, on the, on the transaction log files.  
So on the new replica DB?  
Or did you create a new replica DB Karthik?  
Or it is a it is just a cloned oh oh a new DB?  
Or what is the kind of ohh.

 **Jubaed Prince** 18:28  
Can can someone show the new DB if it's handy that that was demoed yesterday?

 **Karthik Raghu** 18:34  
Yeah.  
I can show it, I can show it.  
But then I think I can show it, but I think it makes more sense for exponential representation, right?  
So do you think that it work like it would be better that we are one second, let me log into the console.  
I can show it, but then again I think I can give a schematical representation of what I'm talking.

 **Ashok KJ** 19:02  
OK.  
Yeah, yeah.  
Karthick and and Prince.  
Why?  
While we are taking up that just a few more just for my better understanding.  
So this is a either it's a clone DB or a replica DB.

 **Karthik Raghu** 19:13  
It's.

 **Ashok KJ** 19:15  
The the DEBEZIUM configuration is done on the.  
If I'm not wrong, I think the transaction log like a bin log file and all the changes including the schema changes will anyway come.  
So I think I saw some questions from Prince of a document like how the alter table or a add table or table one of the column name rename or data type rename on the all those things are going to be handled, but as far as well, uh the museum is concerned.  
I think it sends all the all the information including the schema changes, but I think we are to handle it bit bit differently when it comes to the the the ETL process that happens after that.  
So it when it comes from the from the debezium, I think all these individual table level details are pushed into the Kafka topics for each each table there will be a separate topic in this instance.  
If I'm not wrong and we're going to have everything in there and we need, yeah, sync connector kind of ohh option that we have implemented in, in, in in our previous implementation where we push that data as as a JSON file to do an ADLS or or an S3 so that JSON will contain all these changes right?  
So all the inserts, updates, deletes and schema changes, everything will be coming as part of the JSON.  
I think we can have Avro also as an option, but I think we preferred JSON because it it has a structure and it gives more information visibly that can be seen.  
So Ohh did I quote all the statements right?  
Karthik and Prince is that the plan that we have reached this point?  
The output is in JSON and we need to think about how are we going to integrate or in just this data from this JSON into this new report DB on top of it.  
It's not only just syncing this data, we wanted to have the something like a data marked that may need to be built on top of this data like in your ETL process that we need.  
That will get very closer to the reporting needs.  
So if I'm adding Priyanka and Rupali if you can help me here.  
If I'm understanding this right, so when we looked at the the application sequels, every sequel is director, every column is having, well, uh, every column is having a separate sequel and that is hitting the DB.

 **Speaker 1** 21:44  
Hmm.

 **Ashok KJ** 21:51  
Uh, uh.  
Just just want to understand that part.  
Also, I think that was one of the point I saw in Prince document as well.  
Any any new options that we want to look right?  
I think just want to bring that point also.  
The reason I'm bringing that is like if at all we are building this process from this JSON into this new DB, I think we have to have that also looked into only then I think the Data Mart layer on top of the report DB will be really useful.  
I'm I'm thinking just just stop.  
Correct me to Team if I stated anything differently.

 **Speaker 1** 22:29  
The Ashok I think that is a very right assumption in the old one.  
And if the SQL that we went through in the Excel sheet right, they are from the current application, how did the current application is fetching it right on the reporting layer, which is also in the gold layer, right?  
We can actually, we did that also and we can also do it going forward also that way optimize how the reports are coming, just the logic that we have to make sure we are In Sync with the inference that the you know the report result is correct logic is correct that result am I stating correct to him Prince.

 **Tuhin** 23:02  
Yeah.  
Yeah, yeah.  
And how to get back to Ashok, right, Ashok point right.  
I think you know and I'll just try to explain it in my way like may not be purely the technical terms, but I should right.  
Obviously in some reports, right, you know, they're like on a high level, right?  
You know, there's like, you know, we can.  
Incident scenarios we can like, you know, in the gold layer or whatever in the the my SQL eight that we are creating right we can denormalize the data and like you know take some of the burden out of the SQL join as much as it's makes sense right.  
We may not be able to do it umm all the way just from an efficiency point of view, right?  
But then you know any roll up any calculations that we can do essentially.  
That's what I'm thinking that that that's what that middle layer does, right once the Kafka events comes in, right, you know, it's essentially doing some ETL as you are mentioning and essentially pushing out to some reporting tables.

 **Ashok KJ** 24:04  
Exactly.

 **Tuhin** 24:14  
Then the idea being or.  
Then, like let's say instead of on a very high level, instead of joining 15 tables right, I have denormalized some table as much as I could based on the and, you know, wherever applicable and you know we have to work through that data model together.  
We have done some aggregation wherever it's appropriate, but it may not be the final level of aggregation based on what the reports, features and functionalities implies, right?  
But the idea is that that even maybe that it has rolled up a little bit then instead of like the way I kind of think about it from business logic point of view, right, if I am hitting that database that is on the application tables, right?  
Maybe I'm traversing through?  
I don't know 20,000 records to come back with like 1 record of aggregated pro, but in that my sequel, maybe we can get it down to, I don't know, 100 record, right?  
And that is the level of aggregation we can achieve without compromising on the existing capabilities that the reports allows us.  
And that's essentially what we store right on the on the MySQL late the new MySQL late thing.  
Yeah, go ahead.  
Sorry.  
Yeah, yeah.

 **Speaker 1** 25:35  
So do you go ahead and shop first?

 **Ashok KJ** 25:35  
Yeah, just.  
Yeah.  
Just one one quick update on this.  
So it to him, that's exactly I was referring as data monthly.  
And so where the the denormalized, I think Priyanka you can add this update if I'm saying this differently, but that's the as part of the POC.

 **Tuhin** 25:44  
Yeah.

 **Ashok KJ** 25:54  
I think the the goal layer that was built on on Databricks also was very close to what we are discussing here to him, so it has this aggregated data to a level we cannot aggregate everything.

 **Tuhin** 26:01  
Yep.

 **Speaker 1** 26:02  
Hmm.

 **Ashok KJ** 26:07  
Yes, you said I think some very few of them have may have to be done directly on the, on the on the report, but most of the things we wanted to have the joins and what what is the granularity of the data that we want to maintain.  
I think that is where the the the the biggest work that will go because each report may have a different granularity of data.

 **Tuhin** 26:20  
Yep.

 **Ashok KJ** 26:27  
This data must layer or this table that will be on the reporting DB will serve that purpose to him.

 **Tuhin** 26:28  
Yep.

 **Ashok KJ** 26:33  
That's that's, that's what I mentioned as Data Mart layer just doesn't.

 **Tuhin** 26:36  
Perfect.  
Yeah.  
Yeah, sounds good. Yeah.

 **Speaker 1** 26:38  
To I think, uh, we are calling the the same thing.

 **Ashok KJ** 26:39  
Yep.

 **Speaker 1** 26:42  
Ping DV or data Mart.  
That is the, you know, the just is the same, correct me if I'm, uh, I don't know.

 **Tuhin** 26:47  
Yep.

 **Speaker 1** 26:49  
So one thing when I wanted to ask like yesterday also you and Prince mentioned about normalization and the diagram that you showed right?  
So normalization is on the application DB that you're thinking that you're going out with like how much you can normalize on the DB side, right, remodeling.

 **Tuhin** 26:57  
Yeah.

 **Speaker 1** 27:06  
Or is it on the reporting DB that you are thinking?

 **Tuhin** 27:07  
Yeah.  
On the reporting DB, not on the application DB.

 **Speaker 1** 27:12  
Not on the application, Davey.

 **Tuhin** 27:12  
We don't want to, we we don't want to change anything on the application, baby, right? Yeah.

 **Speaker 1** 27:13  
OK.  
That that. Yeah.  
OK, that that makes sense so.

 **Ashok KJ** 27:18  
Yeah, because yeah, because the the application DB will be mainly built for uh both formance based on the rights.

 **Speaker 1** 27:28  
Hmm.

 **Ashok KJ** 27:28  
That's fully normalized.  
So because that's when the the rights will be faster.  
But when it comes to the reporting DB, I think we have to make the reads faster.  
That's where I think the Denormalization will come into play.  
If I'm not wrong.

 **Speaker 1** 27:46  
And and that's.

 **Jubaed Prince** 27:46  
I'm trying to show something healthy.

 **Speaker 1** 27:47  
That's what I'm trying to.

 **Jubaed Prince** 27:51  
Discussion on guys.  
I just wanna put something also in the equation.  
I was trying to share the share a screenshot but for some reason I cannot send an image in the chat so I'll quickly share my screen.  
Uh again, just to add some topic in the mix.  
So imagine something.  
Imagine this as a pseudo code which does the process event and I think once we get like some event from Kafka insert event we have to do XYZ update event.

 **Speaker 1** 28:25  
Well.

 **Jubaed Prince** 28:26  
We have to do XY and Z and this way we'll have to flow the logic that is going to react to the CDC changes and initially there is a process of taking all the historical data and putting it on the report DB and that report TV already contains that data.  
That's the exercise Rajesh has already done, so this is this is where the contribution.  
Comes in of the CDC and transforming the changes into some sort of aggregated table that that can be used to generate report with simple joins and stuff like those.

 **Tuhin** 29:09  
Yeah, and only thing before we move on.  
Right.  
Because I think it's important to mention.  
So we talked about kind of lambdas as as a concept, right, Priyanka to do do that ETL right?

 **Speaker 1** 29:23  
Yes.

 **Tuhin** 29:26  
So there is two things just we have to kind of keep in mind.  
One is the way it has been at least presented, and please correct me if I'm wrong.  
Right.  
You know Lambda like as a numbers mentioning like it just it app or we can write the called in a in a variety of tooling right in or technology that we want to choose.  
So two things that I think we have to keep in mind, one and I think here Rajesh and Sachin's guidance would be helpful, right?  
It's something that our team will be longer term comfortable with, right?  
Whether we pick a, you know, a framework that's, you know, uh, JavaScript is PHP based or or whatever or like, you know Python, right?  
I don't know because we have really is a variety show of choice and I think that will be dictated by, you know I think what Rajesh and Sachin things is the best fit for us, right.  
I think we should try that right.  
You know, based on the teams expertise there and the second question this is came from Matt, right, the idea is that even if we are starting with a Lambda approach and serverless approach initially.  
Priyanka, one of the things that made is is cautious about right, you know it is consumption based, right?  
You know, it's probably not as expensive as databricks it here, but you know, eventually I think we have opportunity to, umm, you know, move this into as a like a, I don't know, let's just use the word mini application, right.  
And it can be hosted as its own microservice.  
If we wanted to or that just does the work and you know we can put it on dedicated, you know machines.  
But is there a way we could ballpark depending on the data volume that we see?  
Like you know how much the cost is going to be for the Lambda functions to run right maths a little bit high.  
Matt is a little bit hypersensitive about that based on the, you know, some of the pitfalls we had with data bricks.  
So I just wanted to mention that as well.  
I'm not sure if I explained it correctly, I'll be more than happy to try to explain further, yeah.

 **Speaker 1** 31:49  
No, I I think that was a good explanation for me.  
I will try to get back to you with the approximate one and looking at the data and all by next week, right?

 **Tuhin** 31:58  
Yep.

 **Speaker 1** 32:00  
But I'm not sure how close it will be because you know but.

 **Tuhin** 32:00  
Yeah.  
Umm, sure, yeah.

 **Speaker 1** 32:05  
But yes, I think ballpark, you wanna know how much like consuming it and then running the Kafka as well you want to include right and you know on Lambda function and then the reporting DB itself as well or just the middle part because we are keeping the reporting DB we are keeping the application, it's the middle consumption process that you're saying how much it will cost for Lambda function correct.

 **Tuhin** 32:05  
Yeah, ballpark.  
Yeah.  
Yeah.  
Karthik what what is the bigger worry variable for you?  
Obviously the reporting DB is gonna have its own cost.

 **Karthik Raghu** 32:31  
Umm.

 **Tuhin** 32:32  
Like you know, what do you think is important for you to?  
Because the question is gonna come to you at the end of the day, right? Yeah.

 **Karthik Raghu** 32:38  
Yeah.  
So I think doing like, you know Lambda functions are not that costly, right?  
So I think it's very cheap like it's.

 **Speaker 1** 32:44  
Correct.

 **Karthik Raghu** 32:47  
I think it is in many cases.

 **Jubaed Prince** 32:48  
I don't know.

 **Karthik Raghu** 32:49  
I think it's much cheaper than having an easy to instance.  
Uh, but only thing is like we cannot store any data like we have to have any additional data store that we can that that we have to make like that.  
What I mean to say is the data is not oops.  
Data is volatile.  
Basically like you know, it's it's just invisible code.

 **Ashok KJ** 33:10  
People to start different, yeah.

 **Tuhin** 33:11  
Yep.

 **Karthik Raghu** 33:12  
It just runs the code right?  
So I would say like you know the middle layer can be on a machine.

 **Tuhin** 33:15  
Yep.

 **Karthik Raghu** 33:21  
Umm.  
Or an easy to instance, or maybe a service which will which will have this data like which host this data and then the the part that catches the data from Kafka and then pushes it to my SQL right?  
That can be a Lambda function.  
Ohm, we could calculate like how how like we could actually calculate.  
Like how much cost it takes, but I think it is negligible.  
It is what like based on my previous experience, I could say that the cards cost for the Lambda functions since & of negligible uh, it is not that costly as you know you so.

 **Tuhin** 33:55  
Yep.  
Yeah.  
Yeah, and that's, that's what.

 **Speaker 1** 34:00  
That is correct.

 **Tuhin** 34:02  
Yeah, that's what we thought too.  
But you know Karthik, you understand like where the questions coming from, Matt, right?

 **Karthik Raghu** 34:06  
Yeah, yeah, I know.  
Correct.

 **Tuhin** 34:08  
Yeah.

 **Karthik Raghu** 34:08  
Correct.

 **Tuhin** 34:08  
Yeah.

 **Karthik Raghu** 34:08  
Yeah.

 **Tuhin** 34:08  
So yeah.

 **Karthik Raghu** 34:09  
So I think we just have to make sure that we provide some numbers which are comfortable, so that Mac and Mac can feel confident about it.  
So that's what I was thinking.

 **Tuhin** 34:19  
Yep.

 **Speaker 1** 34:19  
The one basic question I'll ask, which I wanted to ask earlier also where I I'm assuming the Kafka is installed because you mentioned a Visio is installed.  
Also, where is it installed?  
Is it like AWS only? Yeah.

 **Karthik Raghu** 34:32  
Yeah.  
Can you?  
Can you share?  
I share the screen.  
I shared the screen like you know this is just I was while we were discussing it.

 **Speaker 1** 34:35  
Umm.

 **Karthik Raghu** 34:39  
Just do it all.

 **Speaker 1** 34:39  
Hmm.

 **Karthik Raghu** 34:41  
So this is the source database right?  
So this is the source data is coming from.

 **Speaker 1** 34:43  
Hmm.

 **Karthik Raghu** 34:49  
So let's assume like this is the production database first cluster.

 **Speaker 1** 34:53  
OK. Hmm.

 **Karthik Raghu** 34:53  
I mean to say, OK, so we have created a clone and then we have created this cluster.  
So this is just a clone.  
Whatever we have here, uh, it's just a replica.

 **Speaker 1** 35:04  
OK.

 **Karthik Raghu** 35:06  
It's not a replication DB, but it's just a replica, so at that point in time, whatever we have would be cloned into this one.  
This does not have any kind of CDC information, if any new data comes in to the existing production, nothing will be there here.  
This is for the POC I'm talking about, right?  
So it has nothing to do with the final this.

 **Speaker 1** 35:23  
OK.

 **Karthik Raghu** 35:27  
So this is the clone cluster that we have.  
OK, so let's assume that.  
So this is the source database.  
It could be the clone Custer or like you know, I I'm coming back to the vision that we have, the thought that we have for the project, but before that, like, let's assume that this is the source, my sequel.  
Right.  
And then we have both of these things.  
Ohh.  
What?  
What of these things already there in?  
I I need to see 2 mission like you know the we have installed Kafka plus they're busy in connect on an EC2 machine and then this is there on the same VPC where this source database is there so this same.  
The same environment.

 **Speaker 1** 36:18  
OK.

 **Karthik Raghu** 36:19  
OK so.

 **Speaker 1** 36:20  
So Karthik, one question I know I'm interrupting on the my SQL you said the the clone DB right?  
I got done Data so does it have the current prod data or is it like we have to populate it because earlier Prince was talking about populating the data so that is also I wanted to get it reporting DB.

 **Karthik Raghu** 36:40  
No, don't get confused.

 **Speaker 1** 36:40  
We'll talk later, but the OK.

 **Karthik Raghu** 36:41  
Just just let's go step by step.  
Right.  
You're you're getting confused again.

 **Speaker 1** 36:44  
OK.  
Go ahead, go.

 **Karthik Raghu** 36:45  
OK, so so let me complete it then I think so for the POC, what we did is we don't we we did not want to mess around with the production database, right.

 **Speaker 1** 36:45  
OK, go ahead.

 **Tuhin** 36:46  
I.

 **Speaker 1** 36:51  
Chief.

 **Karthik Raghu** 36:55  
So yesterday or yesterday what we did is we took a clone of this production database and we created a testing cluster.

 **Speaker 1** 36:55  
OK.

 **Karthik Raghu** 37:01  
So at that point in time, we have whatever is there in production, so yesterday's it'll have the yesterday's data until yesterday's data in this trust them.  
OK, so this is just for the POC, but actually when it comes to the implementation, right, so it will be connected to the live cluster wherein we will be able to get the CDC Reformation as well all the new data will be coming in and that is where the seriously comes in the picture.

 **Speaker 1** 37:28  
Umm.

 **Karthik Raghu** 37:29  
But we did not for this proof of concept like initially to start with, we did not want to mess around with the production database, so we have to we have to choose one more clone database.  
That's so that we don't actually mess around this live production, right?  
So we are experimenting with this with this clone.  
Even if we want to say we want to capture, check and see if we can capture the CDC, what we would do is we will simply write some information into the database and then see if that information is being captured by the CDC process.  
So that is for just for the proof of concept, but this is source right?  
So we have these two installed already on a machine.  
It is actually being connecting to the like.  
You know my sequel server, so in this case like in the POC case, it is connecting to the clone cluster.  
Remember this is a clone, not a replica.  
Replica is like you know it.

 **Speaker 1** 38:27  
Yes.

 **Karthik Raghu** 38:28  
It kind of has all the new information coming in, right, so this is not a replica server, it's this is just a clone server.  
OK.  
So it's just like taking a backup of this cluster and then restoring it to a new cluster.

 **Speaker 1** 38:39  
OK.

 **Karthik Raghu** 38:42  
OK.

 **Speaker 1** 38:43  
Yes.

 **Karthik Raghu** 38:45  
Yeah.  
So, so so this is the idea, right?  
So this is already done.  
Like you know, our like we are able to connect with the DEBEZIUM connector into the my sequel and then the data is coming and the change data is coming into Kafka.  
So we have to find a way to transform this data, whatever is there in Kafka.  
As of now, we did not filter any tables because we did not know what the reporting tables are.  
But I think people here very well know that we can filter the tables like what debezium have to that deep division has to fetch the data from right?  
Like it then we can filter all the tables so that we want to be, you know, taken into uh Kaushika like or maybe ingested into Kafka.  
So whatever data is there in the Kafka, I think this is the application layer that we are talking about which will transform whatever data is there like.  
This will have the raw data, right?  
So you just mention like specify debezium that OK, go to my SQL and get all the change information from all these tables like I can specify filter and then get all these data from all these tables and then put it into Costa.  
So we need a way to kind of read the data from Kafka and then like pick whatever fields that we want, arrange them in such a way that it is easier for the reporting front end to fix that from the my sequel.  
So after the data is transformed, we will just push it to my SQL.  
So that is where I was talking about.  
Like whether you want to use a sync connector from Kafka or to my SQL if we use it.  
I don't know how the data can be transformed.  
Into my sequel.  
Like you know, there are multiple there.  
There is some data coming from multiple tables right?  
So I can't have multiple tables in my destination my sequel.  
I just need to have one table which will have all the consolidated information transformed information.

 **Speaker 1** 40:40  
Umm.

 **Karthik Raghu** 40:46  
So you further I think it is just a very what you say ohm very straightforward configuration for my team to be able to fix the data.  
They can write code and then they can fix the data from this table and then the user will be able to download the report whatever report that they have.  
So in this case, what is happening currently?  
What is happening is like, you know, before databricks right, it was used to just get the data directly from this reporting front end used to fix the data directly from my SQL source.  
So since the data is scattered across, you know all the different different tables or whatnot.  
Uh, you know, the process was very late, so that is where we started introducing data bricks and then say all the ingest all the data into databricks.  
And this was again connected this reporting front end was connecting to databricks APIs to get that information.  
So since that has not worked out well, at least So what we want to do is we want to implement this new approach wherein the reporting front end will not connect to the source database, but in indeed it will connect to the new destination database which will have the data ready like you know it's easily accessible data or it's need not query at the entire database to get the uh information.  
It no joint queries, no nothing, right?  
So it's just a simple select query and then it will yeah.  
Ohm fits all the information and then the user will be able to download the report.  
This layer I think we are talking about the the transformation is what we are talking about.  
Maybe we may use Lambda here?  
I don't know that you guys have to recommend.  
That is why you know, this is what our thought process is, but I don't know.  
Like what you want to recommend that I'm not sure.  
Uh, you know this.  
Whether this layer could be the Lambda layer that you want to do or like you know, So what this this color layer right?  
So is this a Lambda or we can give an EC2 as well here and then this one?  
Also, we need a sustainable solution wherein like you know, if this server goes down right, this is an easy to install.  
This can go down anytime, so I don't know what happened.  
It's gonna happen if this goes down.  
Like let's say I want to restart this machine.  
What happens then?

 **Speaker 1** 43:14  
Umm.

 **Karthik Raghu** 43:15  
I want to restart this machine because this is capturing the CDC information, so I don't know what is the outcome if this server gets restarted because this is not a managed service right?  
This is this is something that we manage.  
So I don't know whether this is sustainable, but you know, I'm just pulling it out.  
This is the overall thing that we are planning, so maybe prints are doing.  
You can add uh.  
If if I missed anything or do you want to add up anything here?

 **Tuhin** 43:50  
No, I think I think that was a good overall and thank you for the thank you for joining the diagram.  
Priyanka, I think you think a Ashok weather line right up to this point that's that's that's essentially what we've been talking about, right, so OK.

 **Speaker 1** 44:02  
Yes.

 **Tuhin** 44:09  
Uh, OK, then I think the question is like, you know, how are you gonna write the transformation then then the technology selection they are that is maintainable.

 **Karthik Raghu** 44:16  
Yeah.

 **Tuhin** 44:20  
And I think are the questions valid there too in terms of, uh, kind of failover and redundancy, I don't know.  
Do you like the like?  
You know that that we can talk about because that question is there then kind of then we can move towards the my sequel discussion.  
Maybe we take the points report because we have a Sachin and Rajesh.  
Here too, we can, you know, start to, you know, start doing some modeling in terms of what we think As for the points report, right, what are some of the tables that may have to be kept created maybe that gives us a point of reference.  
Prince, you have your hands up.

 **Jubaed Prince** 45:00  
Umm.  
Yeah.  
So I I want to understand from Ashok and I I want to say that the discussions we are having so far is really good.  
But there are some tiny adjustment that are required in some of the concepts, but I don't think pointing out those at this moment is really important on a high level.  
I feel like we are coming to the same page where I still wanna hear from Ashok.  
The approach that you are talking about about the Data Mart and and and the approach that Karthik demonstrated now, do you feel somewhat aligned?

 **Karthik Raghu** 45:30  
Yeah.

 **Jubaed Prince** 45:40  
Do you see any any like big difference that you want to bring to the equation anything fundamental where missing in in this approach?

 **Ashok KJ** 45:50  
Uh, no, no, no, no brains and Karthik.  
So I had the similar approach in my mind and the only difference I I won't say it's a difference, but only different option had in my mind is like will will have a A, have them converted all the streamed data from the Kafka topic will be streamed into a into a JSON and from the JSON we thought of looking at the I I personally thought from the JSON we can do the ETL process but I think here what we are referring is like directly let us pull from that topic itself or something like a like a.

 **Karthik Raghu** 46:01  
Umm.

 **Ashok KJ** 46:26  
Streaming job that we can build or or from directly take it from the Kafka and like do the ETL.  
So on the fly, ETL is what here I I was thinking a bit differently by storing that Jason in in a in a in a storage layer, that is that is the only difference or else whatever is highlighted in a in a in a color by Karthik a think that is that is the place I had this small difference but but otherwise I think everything in line to what we we also thought Prince and Karthik so.

 **Karthik Raghu** 46:50  
The.  
Yeah, that's the question mark.  
Right. Ashok here.

 **Tuhin** 46:59  
Yeah.

 **Karthik Raghu** 46:59  
Like, that's why that's where we put a question mark there and then it is read like basically that, that, that that is the thing that we want to make sure that we don't know like you know that that is open on the table.

 **Tuhin** 47:03  
Yeah.

 **Ashok KJ** 47:10  
Like.  
Sure. Sure.  
Yeah, I I.

 **Karthik Raghu** 47:12  
So that's the reason why I colored it out so that we can just maybe discuss about it and then finalize what would be the right approach.

 **Ashok KJ** 47:14  
But.  
That that's.

 **Karthik Raghu** 47:18  
And then when you say, Jason, right, so do you mean to say that we need a Nosql database?

 **Tuhin** 47:19  
Yeah.

 **Karthik Raghu** 47:23  
And then we push that data into into those this thing or like what what is your suggestion?

 **Speaker 1** 47:25  
Umm.

 **Ashok KJ** 47:27  
No, no, no, no.  
I was speaking thinking like an S3 bucket where we support this this on data as a as a as a files.  
So that's how it was.  
It was a Yep.

 **Karthik Raghu** 47:35  
OK.  
OK.

 **Tuhin** 47:37  
So in.

 **Speaker 1** 47:38  
It was actually also thinking to go ahead please.

 **Karthik Raghu** 47:38  
OK. But.  
Umm.

 **Ashok KJ** 47:41  
I I had couple of more points if if OK. Sorry.

 **Tuhin** 47:43  
No, I.  
No, no, actually great.  
You know, I just want to talk through the because you know, it's a it's a good concept and there is a reason you are bringing this up, right?

 **Ashok KJ** 47:55  
Correct.

 **Tuhin** 47:56  
Can you kind of talk through your process in terms of like you know what kind Karthik is saying and why file system maybe a better approach or like you know if you could talk through that a little bit please? Yeah.

 **Ashok KJ** 47:58  
Yes. Yeah.  
Yeah.  
Yeah, that's the sure.  
Sure.  
Then that's.  
That's the point I was about to come.  
Those are the two points I want to talk the one about what if some of this data being not transformed real time and there is some downtime.

 **Tuhin** 48:10  
Yeah.

 **Ashok KJ** 48:18  
So we have the files so that is one reason I wanted to have it in the in the in the S3.

 **Tuhin** 48:20  
Umm.

 **Ashok KJ** 48:24  
That is one reason and the second point is about what if the the Debezium and Kafka had to be restarted or or something gone wrong and things are not not working right.  
I think that is when whatever the first time load process that we are building to push the history.  
I think a similar one we have to do based on the snapshot.  
So once we have all this data in in the storage layer, we can maintain the snapshot of.  
What is the last snapshot of data that we received into the ADLS and from that point on, because we can go back to the the my SQL storage and take the incremental like like 6-2 days of downtime or say one hour of time down time right?  
That particular snapshot will be really helpful.  
That's that's the reason I was thinking we don't know what because this is not completely managed, not a managed service service that we are using.  
It's going to be maintained by ourselves directly, so that's where I thought maybe a file storage will give that advantage.  
I know there is a small cost, but you if I'm not wrong, just threes is comparatively cheaper.

 **Karthik Raghu** 49:32  
That's fine.  
That's fine.  
I'm not worried about this three cost.  
I think it's very cheaper.

 **Ashok KJ** 49:35  
Yes, yes.

 **Karthik Raghu** 49:36  
Ohh and you know.

 **Ashok KJ** 49:37  
So they I think we can have that kind of a failover advantage of I'm doing a history load for that missing data or missing one hour missing 2 days.  
I think that can be handled if we have that kind of snapshot in our layer because once we have this completely real time, I'm sure I'm not saying that things will go down very frequently, but whenever this goes down, I think that we have that additional thing.  
But there is there can be an alternate on the topic I'm talking about.  
Also like what we can think of is like directly go to the report DB and we can get the Max updated time by us that also can be used for the incremental change.  
I'm not.  
I'm not denying that, but this may give some more details in terms of what we received.  
Like if you want to do a backtracking of what really happened, I think we have the JSON files too.  
Check. Check.  
That's the point.

 **Karthik Raghu** 50:27  
So how are you going to pull?  
Like you know, I know you talked about like transforming it into they just on and then accessing it, right?  
So like here in between like you know there has to be something which will do that process right?  
So what do you call that?  
Like what?  
Like, well, what makes it possible to do?

 **Ashok KJ** 50:46  
Yeah, let's beaming engine.  
I think cutting pre income may be able to give more details.  
I think the the the the Spark streaming is an option that we can pick and that can do both the things right in terms of converting or transformations that we want wanted to do to the Data Mart layer.  
That can be an option, and that can be other other options also, because what how we want to do this transformation right?  
That depends.  
So we can, but I think Greene got me have more more details on that.  
So we wanted to go back and get back to you, Karthik, but on a high, but there are spark streaming is an option and we can have a different like in, in, in up AWS World.  
Or something like that that can continuously fetch all this streaming data into into JSON.  
I think that is also an option that we can think of, but I don't want to bring in new tools.  
That's why I thought of Spark Streaming is an option.

 **Karthik Raghu** 51:41  
The whatever we do, I think we need to keep a keep an eye on the on the cost, right.

 **Ashok KJ** 51:44  
Stop.  
Correct, correct. But.

 **Karthik Raghu** 51:46  
So, you know like like I'm just worried, like, you know, basically if we introduce kinesis like, you know what?  
What's the what's?  
What's the cost of gonna be?  
Right.  
So that is that is where I'm thinking about.

 **Ashok KJ** 51:57  
Yeah, exactly.

 **Karthik Raghu** 51:59  
Yeah, but that's fine.

 **Ashok KJ** 52:00  
That's why I didn't.

 **Karthik Raghu** 52:01  
I think maybe Prince, do you have any questions or to him like for for the recommendations?

 **Jubaed Prince** 52:07  
Yeah, yeah.  
I have a couple of things that I want to discuss on one thing is Umm Ashok's point on having some sort of database to capture the event streams and then doing the processing on that database keeps us safe from losing stuff when we are.  
Trying to do a real time processing basically so it it gives us certain delay but it also helps us like be able to go back and do some batch processing there.  
I'm Ashok.  
I I don't know if this is the right point to discuss technology, cause I I know of course everyone will have to go back and homework but but these are good technologies, but I also want to talk through about umm is it?  
Is it a good idea to have S3 on as a layer like a file based layer to capture all this data?  
Because these are gonna be real time data, so a lot of heat are gonna be concurrently happening.  
Does it make sense for us to capture those in in a like a database format?  
I'm not sure.  
Umm, I again also capturing those and processing them versus real time, processing them through some sort of transformation layer that has been exposed here.  
I I think we should go back and figure out like, what is the minimum tooling we need to form the MVP or a POC around that.  
While we do that, we'll learn more and we can optimize that.  
So that's one and #2 is.  
Yeah.  
Umm.  
Matt is asking a question about Lambda scalability and I think I'll I'll take a stab at that answer and then I to answer that as well.  
And Karthik is here too.  
But Lambda usual what it does is it's auto scalable so as as the load increases it it increases with it.  
But with Lambda we gotta be careful that we we don't overuse it or we don't misuse it.  
So there is some sort of optimization is necessary when we're doing code here, because it's easy to get high those if we are not careful.  
But in in a POC stage I think Lambda is a good option because because it gives us a quick easy way to kind of validate our concept and later on we can probably have some sort of microservice or EC2 or whatnot we can really figure out OK what is the best way to host this.  
But for that time being, I think to keep things simple, we can think about just what is the application inside that transformation rather than how it is kind of, uh, hosted or in terms of infrastructure, I think we can start thinking a little bit later.  
From this point on.  
But yeah, Priyanka, you have your hand raised.

 **Speaker 1** 55:36  
Yeah, actually, uh Prince.  
Adding to your point in Ashok, I actually find that a good you know.  
Could approach you take to land all the uh, you know files and to a storage it's it's not too much cost also but it will also that point of failure will stop from there right?  
Otherwise, it if everything falls, we'll have no updates and we won't be able to track back and I think Karthik, to your point when you were saying how it'll land there, one option is we can configure the Kafka connector to learn make the S3 bucket as the sync source right sync of the Kafka.  
So whatever it takes from there and then it can, you know, just landed inside.  
So whenever there's a event, it'll just take the CDC files and land it into the S3 bucket on the Lambda side or whatever we are taking after that in the process, we can trigger, we can have an event trigger.  
There is an event right that we can trigger on on match point of Lambda.  
Actually, Prince, we were also discussing that yesterday that there there is there are certain limitations on you know there are certain limitations on Lambda that is like for example it has some memory limitation which is memory memory.  
When I don't see a problem on at least on an xcloud side, because it goes to 10 GB for one, you know one Lambda function and that is not a problem on our side, but well, there is like concurrent execution.  
How many of them?  
That is also a limitation, but we can have that increase by W support.  
Still, there is some person per account per region.  
There is limitation, so we have to make sure that we are accounting for that code.  
Latency does not come for your you know your problem because it's like milliseconds.  
That should be alright for us, and I shall feel free to jump in in that one if I'm missing something.  
But the one thing I agree, Prince, we can use it for the POC because the same code can actually be hosted on containers and there are two options that we can go for instead of Lambda if we think our processing is higher as ECS and have or having a fargate, I would not suggest fargate in our case because it's again you know it's cold start right it comes up and goes down, yeah.  
So yeah, I was thinking ESG having auto scaling, if we have are going, if anything we think that Lambda is not fulfilling our function, we can host the same code on the case.

 **Jubaed Prince** 58:08  
Like.

 **Speaker 1** 58:08  
And I think that is what Karthik you were talking about having EC2 instances and managed by ECS, right?  
Yeah, please go ahead.

 **Karthik Raghu** 58:17  
My question is like you know, just remember whatever whatever solution that we are trying to provide, right must work for Azure as well as AWS.  
So if it is just Lambda then I think what do we do with the Azure?

 **Speaker 1** 58:30  
Yes, umm.

 **Karthik Raghu** 58:33  
Like we can't use Lambda for Azure right?  
So I think we have to find something similar which is like Azure functions, uh, to be able to do that, then the code will again be changed, right?

 **Speaker 1** 58:37  
So.

 **Ashok KJ** 58:38  
That is, you can so.

 **Speaker 1** 58:40  
Does eventually.  
Yeah, go ahead Ashok.

 **Ashok KJ** 58:43  
Yeah.  
So you're few points, I'll just try to remember all of them.  
So I think for one of the the reason one is from Karthik on what is the equivalent?  
I think we have Azure functions and we have a WWW but equivalent AWS Lambda equivalent.  
We can use Azure functions there Karthik and few more points.  
I just want to go back for Prince question on do we need to hold it as a JSON or is it OK to be in in table to maintain this it it depends on what is the transformation plan, the logic that we want to build.  
So if we are going to build everything like a sequel way like a procedures and those kind of things, yeah, maybe a table way is better, Prince.  
But if we want to do it in in different like a different detail process for the transformation, I think we can go for files or tables.  
Doesn't the that that that may not make make a big difference there.  
Then the second part is is the.  
I thought what other item move the questions so in terms of POC, definitely Lambda is an option but like Priyanka told I think we can look at when when we want to go for the actual implementation.  
I think we can look for look for the different different options also like what Priyanka suggested and in in in this particular case I think Azure function is an equivalent Karthik.  
So hope I answered all three three outstanding questions.

 **Karthik Raghu** 1:00:10  
Yeah, I know, I know.  
I know we can do it with Azure functions, right?  
So it's just that, you know, I'm just trying to find a solution which works both the ways right?  
So see if you like.

 **Ashok KJ** 1:00:19  
Yes.

 **Speaker 1** 1:00:20  
He.

 **Karthik Raghu** 1:00:21  
I'll, I'll tell you.  
Right.  
Like Azure is our.  
Umm, like strategy moving forward, right?  
Like it's basically we would be moving all our existing clients to we will, we might have some plans moving or existing clients to Azure platform.

 **Speaker 1** 1:00:35  
Hmm.

 **Karthik Raghu** 1:00:39  
So this is all for our existing customers like AWS just, yes, existing customers, but maybe like we are getting very bigger customers into onboarded into Azure by the end of like not enough like better middle mid of this.  
Yeah, I think you have very big customer of landing on our Azure platform.  
So whatever solution that we give must be aligned to the fact that it should work both the ways, right?  
So let's say we spend some effort on on on this Lambda thing, right?  
And then you keep we have to go there and then go back and then reimagine if something for Azure.  
Then I think it will be a problem like.  
Basically, that'll take more time since we are trying to implement the two different technologies.  
We have to think that like instead of land or let's assume that I have a server like an easy to server which will actually take care of all these things right?

 **Ashok KJ** 1:01:28  
OK.

 **Karthik Raghu** 1:01:36  
So in that case it's a completely different what you say approach.  
I could, I could say that OK, if it's an easy to on the AWS then it will be am on Azure.  
I'm not saying we have to stick to that plan, but basically we have to also things that we don't have much time to do this.  
Then if you do Lambda for POC, uh and then we figure out that OK, if we have to get this working on Azure then we have to work few more days to make this compatible with the Azure functions and then do it right.  
So that's that's what I'm worried about.

 **Speaker 1** 1:02:15  
So oh, thank you, God.

 **Tuhin** 1:02:15  
So there's a couple of Priyanka.

 **Speaker 1** 1:02:17  
Go ahead then.

 **Tuhin** 1:02:18  
I'll.  
I'll, I'll.  
I'll try to like chime in right and maybe Priyanka you can help me answer those like hopefully that clarifies it right.  
So the question number one, I think the assumption is because code is code at the end of the day. By and large, right?

 **Speaker 1** 1:02:27  
Yeah.  
Yes.

 **Tuhin** 1:02:34  
So if we start with Lambda or Azure function right and you need to convert it into a I don't know what I call it.  
I'll just call it an app, right?  
That gets hosted on a on a computer, or whether that's C2 or like, you know, equivalent in in Azure, right?  
How much do you say?  
I obviously depends, right?  
You know, if we if we build it, if we build that app or not JS or if we build the app and something else right, that's gonna be a little bit different, right.  
But in terms of the transformation code, how much of that is transferable between a serverless function and if we want to move that into, move that into a kind of a you know app that has calling.  
So that's kind of question number one.  
And question #2 is that the exact one that Karthick's bringing up right?  
Is it are we talking about days and weeks?  
Like let's say we have to do cloud functions as well as lambdas in terms of doing doing the POC right?  
How much of an effort that is in Priyanka Nashik if you can kind of try to thoughly estimate that I think maybe that that gives us a path forward right?

 **Speaker 1** 1:03:51  
Hmm.  
So.  
Don't actually through that point I wanted to add an Ashok keep me honest here.

 **Tuhin** 1:04:00  
Yeah.

 **Speaker 1** 1:04:03  
I did so when actually this is a good point that Karthik brought right that we have to make sure that it will.  
It might, and you also brought that up yesterday when we were discussing, right.  
And one thing that we have to keep in mind like this is for us also in you know in next cloud also when we are developing whether it's in Lambda or not or on SQL it's similar like SQL Spark that we are talking about right.  
Although nobody's calling it out, spark or pyspark, but that is the code that is going right.  
That is where I think you are also living to.  
So how right now you have those SQL queries and you can transfer port them into other codes and just have to work on that, right?  
The same kind as spark code will be once we convert it into spark code that will be spark code which can be hosted.  
This is just Lambda function is just a serverless engine that it's being hosted on and we have to.  
That is one thing we have to keep our developing.  
This is for like assault, so I'm calling out that most of the configurations which are for the spark streaming or spark, you know, parallelism that should all be in the code and we should not depend on too much on the configuration side, even if it is, it should be easily, you know like how in databricks it's right now that you can just set the codes and all in the code itself.  
So that should give us that portability moving from one place to another, at least on the logic side, we should not be doing too much work when we do, whether we move it to an open source, whether we move it to on Prem or we move it to functions, not even easy to instance, right.  
Uh, similarly on the Azure side also right.  
And Ashoka you can jump in if I'm going in the wrong direction here.  
Who and that is, that is the I think that is the expectation and that is what our UN, Karthik were also talking about.  
Am I?  
Am I going right to him and I should feel free to jump in there if that is the same thing.  
You were also thinking.

 **Ashok KJ** 1:05:59  
Yeah, I know.  
I was.  
I was typing for a karthick's question. Yes.  
Yeah.  
We have to look at exactly the compatible option because it's about how we execute.  
Is the Lambda or the function, but or the Azure function?  
But what we write should be compatible between both the both the ecosystem.  
So that's that's that's that's definitely should be in our mind.  
We'll we'll get back in in that version as well.  
I think we have to literally see like what we initially planned for Lambda.  
I think is it completely compatible there.  
Ideally it should be, but still just because this is the point of discussion that we can get back.  
Priyanka so.

 **Speaker 1** 1:06:34  
So one thing, Ashok, one thing was coming to my mind like these things like the main problem might be if we start using Boto 3 on that side then it might not be compatible.  
So we have to make sure we don't go to AWS Boto 3 and we keep doing it actual, you know, native pyspark.  
What do you think?

 **Ashok KJ** 1:06:53  
Right, right, right, drink.

 **Tuhin** 1:06:56  
So so I think.

 **Karthik Raghu** 1:06:57  
Yeah, I I think that makes sense, correct.  
Like because I think they themselves are, I think so.  
I don't know what's the future for Moto 3 like.  
Basically, I think I remember.  
I think we had some issues with the few of the modules getting when you say expired or not supported or something like that in the previous experience like yeah whatever it is, I think let's make sure that it is standard across Azure as well as this right?  
So, but I think to a point, I agree to the fact that if we use the functions like, I'm just calling functions basically because it could be Lambda functions or it could be Azure functions right?  
If if I'm kind of leaning towards the scalability aspect.  
That that's one good thing.  
But we may have to do some test, like uh Prince was mentioning, right?  
So I think we have to see like you know we can do some load testing and see what kind of cost impact it creates.  
But yeah, just just whatever you guys do, right like this, consider the fact that we are we are considering, you know we we are putting Azure into.  
But they are into the picture that you have to do whatever you have you propose.  
Keeping in mind that Azure is all is is there in the picture.  
That's all I wanted to say, Prince, and doing it.

 **Jubaed Prince** 1:08:26  
Umm, a quick question, does Lambda support Docker image?

 **Karthik Raghu** 1:08:34  
It does.  
There's something called as Stargate.  
It does, but I think it it's just, you know, it's a serverless, right?  
You don't get to you.  
You you don't get to choose what you want like basically you put the code and then you just pay for the you pay for the pay, pay for the money.  
Uh for?  
For how many hours of execution that you consume?  
So it's just serverless.

 **Jubaed Prince** 1:09:02  
Because here is what I'm thinking.  
If if we do put that application code in a Docker image then it should be easy to just put it anywhere right?  
Be it as your or Edwin S and if if Azure function and Azure function and AWS Lambda also supports that, then it it should be very smoothly be able to transfer them but.

 **Karthik Raghu** 1:09:16  
Yeah, that's a conventional that should be correct.  
Correct it.  
No.  
Lambda is basically an area for you to execute.  
The code prints, so it provides you the required tooling to like you can say like OK, I want to run a Python script.  
Then I think if you give you a way to put that Python script and then you can run it right.  
Everything else is managed by AWS and then let's say your Python script will run for 3-3 minutes, then you will only be charged for three minutes.  
The wedding in conventional infrastructure, what will happen is like, you know, in order to run that code, whether you use it or not, you will be paying for the for the time that like 24 hours you will be paying, right?  
So let's say I have three or four schedules in a day, and then each job will take 3 minutes.  
So just for those four schedules, I would be running an entire virtual machine all the time, 24 hours, but with Lambda function we have that flexibility where like you know with our Python code is just running for three minutes.  
You just pay for three minutes and then if you have four scheduled the day you pay for 12 minutes.  
That's it.  
So we don't have the flexibility to choose.

 **Jubaed Prince** 1:10:38  
Right.  
My question Karthik is around the code base in it and the and the environment that it is running on.

 **Karthik Raghu** 1:10:49  
Yeah.

 **Jubaed Prince** 1:10:49  
Can we put that in a container?  
That that's kind of like the question because then it's portable, right?

 **Karthik Raghu** 1:10:54  
Ohh.  
Yeah, but in a way, I think it is portable like you know it's it's like wherever you run it right, it's there are two options here.  
Things like one option is ohm.  
You explore the options like.  
Let's say they come up with some code, right?  
So that's what Priyanka was mentioning.  
Let's not use Boto 3.  
Have to use something else, right?  
So nothing for a priority today might be using some common things to execute that which can be reused in Azure functions which is equivalent to Lambda in 8 W so since the code is same which works in both the platforms like you know if if they use some spark or whatever right so that code will be same across everywhere.

 **Jubaed Prince** 1:11:32  
So.

 **Karthik Raghu** 1:11:43  
It's just like instead of running it on Lambda, it runs on uh.  
I see.  
But what we need to make sure is we don't use any.

 **Jubaed Prince** 1:11:50  
Right.  
So is there any concern around migrating those thing?  
Like will it take a lot of time, small time?  
What's the thought there?

 **Karthik Raghu** 1:11:58  
Yeah, that's that's something.

 **Jubaed Prince** 1:11:59  
Because like I'm thinking that, yeah, yeah.

 **Tuhin** 1:11:59  
So can I?  
Can I jump in a little bit, guys?  
Right.  
I I love the discussion about the tooling and pros and cons and you know, I think that's great and I think it's an important discussion to have, right.

 **Karthik Raghu** 1:12:03  
Yeah.

 **Tuhin** 1:12:11  
But you know, I think obviously we have two paths, right.  
You know, there's a couple of decision has to be made and I think we are branching out into hypothetical alittle bit.  
It's important.  
It's important to talk about.  
I don't want to minimize friends and Karthik, right?  
But you know, so let's do this right.  
So obviously as part of that experiment, right, you know, I think there's a couple of things we need to do in POC, right.  
Someone has to write that code right?  
Whether we're running it on Lambda in a hosted machine, app dockyard, all of that kind of thing, right, there's pros and cons, like let's go ahead and try it.  
Right.  
And if that's OK, sorry.  
Princeton, Karthik, I I didn't mean to, like cut off the discussion because there in a good flow.

 **Jubaed Prince** 1:12:53  
No, I I think you are right on the point of when yeah cause these are these are things we can figure out over time as well, yes.

 **Karthik Raghu** 1:12:56  
Umm yeah.

 **Tuhin** 1:12:57  
Yeah.  
Yep.  
OK, so can I pivot it back?  
I think we have come back with like I wanted to ask the entire team.  
So do we think we have a high level understanding of the target state like with the Karthick picture that they have shared, a Karthik have shared then what I'm really trying to get at because we have about 45 minutes left, right?  
Let's take a look at the tasks that we need to do next next week, because there's quite a lot of things we need to do.  
Let's just make sure if we can figure out, are we willing to take up this task and if we are, who's going to be responsible for this kind of task?  
If you are missing tasks, we can add them and essentially document it right and obviously then we can we can run through it, we can set deadlines.  
And so we, we don't miss that target next week, right?  
So that's what I'm kind of coming up with, if that's OK with everyone.  
Oh, oh, Karthick pens, pianca, Ashok Rajesh.  
Sachin, if that works.

 **Karthik Raghu** 1:14:08  
What's for me?

 **Speaker 1** 1:14:08  
Yesterday.

 **Jubaed Prince** 1:14:09  
Yeah, yeah, yes.

 **Karthik Raghu** 1:14:09  
Yes.

 **Rajesh Bute** 1:14:10  
Yep, Yep.

 **Karthik Raghu** 1:14:10  
Yeah. OK.

 **Jubaed Prince** 1:14:10  
And I I think we should spend like at least like maybe 10 minutes just going through the door commenting on it and just just making sure we are all able to put everything in it, that our head is thinking of.

 **Ashok KJ** 1:14:12  
Yes.

 **Tuhin** 1:14:17  
Yep.

 **Jubaed Prince** 1:14:24  
And then from there?

 **Tuhin** 1:14:25  
Yep.

 **Jubaed Prince** 1:14:27  
Umm, we can move forward like start maybe we can start commenting, asking questions, raising concerns, stuff like.

 **Tuhin** 1:14:28  
OK.  
Yep.  
Yeah.  
OK, so everyone should have access to this.  
Let me know if someone doesn't.  
I'm more than happy for us to read through it and.  
Umm and adding like you know this, any of the names that we see here, these are suggestions, right?  
If someone really is feeling passionate about it, let's just go ahead and add our name and everything is to do right then.  
If any of the points doesn't make sense, let's make an inline comment for this staging plan.  
So as as Princess mentioned like that, just take 5 minutes and I'll stop sharing.  
And does anyone here not have access to the this talk that I need to add right now?  
I think I added most folks from Kanini as well, but I am double.

 **Karthik Raghu** 1:15:36  
I got in late.

 **Tuhin** 1:15:38  
OK.

 **Karthik Raghu** 1:15:38  
OK.

 **Rajesh Bute** 1:15:39  
I'm able to access it, yeah.

 **Tuhin** 1:15:41  
OK, perfect.  
So maybe we take 5 minutes and or 6 minutes, right and we come back at like 1258.  
Ohh our 12 or 1225 our time right?  
We can stay on the call like, let's just go through it and if we want to volunteer for certain things, let's volunteer.  
If not, if we are missing something, please feel free to add those items.  
Then let's run through that for the remainder of the meeting. Yeah.  
Fancy had your hand up something, sorry.

 **Jubaed Prince** 1:16:13  
No, no, sorry.  
That was by mistake.

 **Tuhin** 1:16:16  
OK, I E.

 **Speaker 1** 1:23:41  
Then I didn't go to party, did not get it.  
The invite for the document.

 **Tuhin** 1:28:35  
Can you see me? Sorry.

 **Rajesh Bute** 1:28:40  
Yes, ma'am.

 **Jubaed Prince** 1:28:40  
Yes.

 **Tuhin** 1:28:41  
OK, so I think now that we spent the time like someone probably should help us drive this.  
Or do we need?  
Do we need additional time like essentially want to run through the staging plan?  
Priyanka.  
Rajesh, Prince new we can drive it, maybe screen share and like go step by step make sure we're all in agreement.

 **Speaker 1** 1:29:05  
Shutter.

 **Tuhin** 1:29:06  
We're not missing anything.

 **Jubaed Prince** 1:29:08  
I'm let's have someone from the community side drive this and then we can pitch into it.

 **Speaker 1** 1:29:08  
Sure.

 **Tuhin** 1:29:14  
OK.

 **Speaker 1** 1:29:17  
OK, I'll go ahead and do that.  
Give me a second.  
So we wanna go through the the document that you shared only and OK.

 **Tuhin** 1:29:28  
Yes.  
Specifically, I think the staging plan for the points report right.  
I think that that section, yeah.

 **Speaker 1** 1:29:35  
Hmm.  
OK, so listen so I think for the report DV you already have Karthik, right?  
Karthik, you'll set up the report DB one question I had is the reporting DB will have.  
The scripts created the the reporting tables because that is also a task we would have correct and as I don't I'm not sure that will be your team.

 **Tuhin** 1:30:03  
Yes, so.

 **Jubaed Prince** 1:30:05  
And Sachin, do you want to add some information about what you did in past two days?

 **Tuhin** 1:30:06  
Yeah, I.

 **Sachin Naikwadi** 1:30:08  
Yeah.

 **Tuhin** 1:30:09  
Yeah.

 **Sachin Naikwadi** 1:30:13  
I am so I think so.  
We tried to create one separate table for interaction report just for visiting purpose we created one single table and we did aggregations in PHP script.

 **Speaker 1** 1:30:17  
Umm.  
Umm.  
Umm.  
Hmm.

 **Sachin Naikwadi** 1:30:32  
OK.

 **Speaker 1** 1:30:32  
Hmm hmm.

 **Sachin Naikwadi** 1:30:32  
And we transferred that data to that newly created table.  
So same thing for .3412 uh by PHP script where going to aggregate the values and it in single table we are going to insert a point support.

 **Speaker 1** 1:30:37  
No. OK.  
Umm.

 **Tuhin** 1:30:47  
Yeah.  
So only only caveat I I do want to add there before like you know all I think that's that's a good thing.

 **Speaker 1** 1:30:51  
Umm.

 **Tuhin** 1:30:56  
Maybe we need to spend some time on the data model.  
Some of the things that we're so do we need to create and I think everyone should that do that together with sachin's guidance because Sachin's probably these subject matter expert we have right in terms of the tables then then we think about what the table is what the functionalities are and do we need one Table 3 Table 5 tables I don't know and what is the level of denormalization and negation we can achieve right.

 **Speaker 1** 1:31:00  
Yes.  
Hmm.  
Umm.  
Umm.

 **Tuhin** 1:31:25  
So maybe we can.

 **Speaker 1** 1:31:26  
Umm.

 **Tuhin** 1:31:26  
That's an exercise that needs to happen, in my opinion.  
I don't know if anyone disagrees.

 **Speaker 1** 1:31:32  
No, actually that is what I was alluding to go to in that we would need that right just and I think Sachin can guide us on that one to have at least for the points report we are focusing on the points report here at least for the points report.  
What kind of data model should be there in the reporting DB on the application side we know all the you know 7 or 8 tables that are to be used.

 **Tuhin** 1:31:47  
Umm.

 **Speaker 1** 1:31:51  
Then it'll be because then it'll be sync with what the Lambda function is doing and what it's aim is, right?

 **Tuhin** 1:31:54  
Umm.

 **Speaker 1** 1:31:58  
What it's calculating where it's going?

 **Tuhin** 1:31:58  
Yeah.

 **Speaker 1** 1:31:59  
Because I think that will be a single exercise, right modeling as well as what are we putting actually there?  
Right.  
The new computation layer, right?  
So should I put that as a task here before we go further initial data migration?

 **Tuhin** 1:32:12  
Yeah.  
Yeah.  
So, uh yeah.  
So let's put it as a test there and and you know, I think all of us can work together.  
I'm happy to pitch in from like functionality point of view as well to keep us on on track, yeah.

 **Rajesh Bute** 1:32:48  
Well, I just want to add to into that one Priyanka.  
So though my sequel 8R days, meaning that it's just blank database currently.

 **Speaker 1** 1:32:52  
Umm.

 **Rajesh Bute** 1:32:56  
OK, so the searching created the one table related to the interaction table.

 **Speaker 1** 1:32:56  
Umm.

 **Rajesh Bute** 1:33:01  
Currently it having the all the columns which we're showing it on the interaction report.

 **Speaker 1** 1:33:03  
Hmm.

 **Rajesh Bute** 1:33:07  
So through the script we have pulling it from the existing RDS and inserting into the MySQL date new table.

 **Speaker 1** 1:33:07  
OK. Yeah.  
Umm.

 **Rajesh Bute** 1:33:16  
So four points we could. Yeah.

 **Speaker 1** 1:33:16  
I think the yeah.  
Yeah, go ahead please.

 **Rajesh Bute** 1:33:18  
So yeah, so for points report, do you have those tables and the queries with you right?

 **Speaker 1** 1:33:26  
Yes, we do have that.  
The excel sheet right that we had initially, yes, we do have that.

 **Rajesh Bute** 1:33:29  
Yeah.

 **Speaker 1** 1:33:38  
And I think, uh, I think that that is what we are aiming for, right, Rajesh.  
And maybe we can have a brainstorming session with everybody there and make sure that everybody In Sync.  
What is going there?  
So the logics are very very clear, although the SQL are there.  
As Ashok also mentioned earlier in the call that we can optimize them as much as possible, right and not have a select query per column kind of thing, right?  
So that will be helpful for us also on creating Lambda function so that we can have optimized one.  
What is going where and all those things?  
Select queries I we can transfer and before going for the session I would actually that we, you know look at the logic and we also come back with what we think is the most optimal approach to put the data.  
So this is I think it's kind of like both kind of thing Data not migration but data transformation logic as well as the debt.  
And I think that is there already in the table. Good.

 **Tuhin** 1:34:32  
Yeah.  
Yeah, I think I think Princesses hands up.  
So I'll just make this one quick comment.

 **Speaker 1** 1:34:38  
Umm hmm.

 **Tuhin** 1:34:38  
Right.  
You know, I was.  
I just doing this back of the envelope calculation with friends, right?

 **Speaker 1** 1:34:43  
Umm.

 **Tuhin** 1:34:43  
Umm, one thing that I think with Sachin and Rajesh has help will be able to do some of the tables that we have to create.  
I think if we are smart in thinking holistically and that's where people who know about the other reports like me and you know, Sachin and like, you know, pitching in, then we don't have to create like you know, we may be able to create a table for example that always needs a snapshot of the user's most up-to-date data, right?

 **Speaker 1** 1:35:01  
Umm.

 **Tuhin** 1:35:13  
Current balance available point.  
What have you that can be used in points report liability report?  
Member report.  
What have you, right?  
So I think there are some optimization opportunities there.  
So, but we can talk through that.  
Just want to mention that then we don't have to like, you know, that will just speed up the work down the road for the rest of the reports.  
Just want to mention that Prince, if you wanna go ahead, yeah.

 **Jubaed Prince** 1:35:37  
Yeah.

 **Speaker 1** 1:35:46  
The umm.

 **Jubaed Prince** 1:35:48  
Kind of worked on and maybe add more to it, but it can be like a bare bone exercise starting baseline position maybe?

 **Speaker 1** 1:35:55  
Umm, I think that is a good idea, Prince.  
So we can have this one working session where we can talk about all the points, but I also want to have a very dedicated session for these tables as a you know, as Tuhin said, that will be helpful and this will to input it very well that actually that is the thought process going for this right that we don't have to rebuild this table.  
Again, this is like this is reusing everything that is possible, so Tuhin and Prince.  
Do you guys want me to move this to the this one and we can have a working session for this before actually we start doing all this?

 **Tuhin** 1:36:31  
Yeah.

 **Speaker 1** 1:36:32  
Maybe we can go through one workshop for this.

 **Jubaed Prince** 1:36:33  
Yes, yes you can move it up, yes.

 **Tuhin** 1:36:33  
Sure.  
Yeah.  
Yeah, let's, let's even have that on Monday morning around this time, so we don't lose that day, right, because Or ohh Monday is Holiday Mondays holiday.

 **Speaker 1** 1:36:41  
Uh doing Monday is it's a holy festival.

 **Jubaed Prince** 1:36:45  
Holly.

 **Tuhin** 1:36:45  
Yes. Yes.

 **Jubaed Prince** 1:36:46  
Holly, come on to him.

 **Speaker 1** 1:36:46  
Holy festival, right.

 **Tuhin** 1:36:47  
Holy.  
Yeah.  
Yes, yes.  
Yeah.  
Yeah.  
So let's do it on Tuesday.

 **Speaker 1** 1:36:49  
I'm sorry, I want to work, but my kid won't let me work.

 **Tuhin** 1:36:51  
Yeah, let's do it on Tuesday.  
Yes.  
No, no, no, no, no.  
I remember you reminded me.  
I forgot.  
So it's my fault.  
So Tuesday, right?  
So let's do it on Tuesday.

 **Speaker 1** 1:36:58  
And up dessert.

 **Tuhin** 1:37:01  
Yeah, yeah.

 **Speaker 1** 1:37:02  
So.  
So I'll put it up here, right?  
And I did actually volunteer Rupali is very good at these documentation.  
We actually went through all this, so I wanted Rupali and Muni to work with whether Tuhin you or Prince and maybe Sachin, right to help you build this documentation and I'll add this as the first pointer.

 **Tuhin** 1:37:15  
Yeah.  
Perfect.

 **Speaker 1** 1:37:26  
So now I think we are all volunteering for it.

 **Tuhin** 1:37:33  
Yeah, the better one.  
Is that what you did?  
Like we just volunteer someone else.  
And that that's a better approach, right?

 **Speaker 1** 1:37:40  
I. No.  
I asked her before volunteering her because she's good at it.

 **Tuhin** 1:37:45  
Yeah, yeah, yes.

 **Speaker 1** 1:37:47  
I see.  
That's that's just how it works.  
Yeah, that does this look OK, Prince, this is what you were suggesting, right?

 **Jubaed Prince** 1:37:59  
Yes, yes.

 **Speaker 1** 1:38:01  
Ohh that thing up.

 **Jubaed Prince** 1:38:02  
And then Add all of your topics as well in this list.  
Basically, it's just like a starting point and then put more in, yeah.

 **Speaker 1** 1:38:09  
Umm to that.  
Actually, I'll put this as two points now because I put it as one.  
So create new tables or create rather discuss because I think.  
Of the new double with the optimal data model this this will be for KANINI suppose to.  
There's at least this clarification on logic for transforming the data on logic side.  
I'll leave that one in.  
Ohh not.  
So I'm I'm just not putting Lambda access, I'm just putting for so that we know bare bone.  
Yeah.  
What is the transformations we need actually to put it, whether it will be an issue, whether it will be in ohh but by shop or anything there's that makes sense, uh.

 **Tuhin** 1:39:13  
Yep.

 **Speaker 1** 1:39:27  
OK.  
Anything else?  
So this is the two things.

 **Tuhin** 1:39:49  
Yeah, and I think.  
CD, EF. Uh.  
Actually all the way up until age is part of that data modeling exercise, I guess.

 **Speaker 1** 1:40:03  
Hmm.

 **Tuhin** 1:40:05  
Then.  
I and Jay to me seems like a little bit of a different exercise, but I you know, that's me, laymen speaking.  
If you go go up a little bit because it's like I and J, right?

 **Speaker 1** 1:40:20  
I to Jay.

 **Tuhin** 1:40:23  
Because now we are getting into like materialized views and things of that nature which my SQL doesn't support the plugins and whatnot.

 **Speaker 1** 1:40:24  
Yeah.  
Does it?

 **Tuhin** 1:40:32  
So the use cases right?

 **Speaker 1** 1:40:33  
Hmm.

 **Tuhin** 1:40:33  
So that's something we would have to think about.  
I can't help there. A Princess.  
That item.  
So like, I don't know.  
Prince, what your thought process was there?

 **Jubaed Prince** 1:40:42  
So test, it's just like when you're thinking about things.  
Just think about this as well and see if it makes sense.

 **Speaker 1** 1:40:47  
Hmm.

 **Jubaed Prince** 1:40:48  
When you open, go under the hood, so it's like your question, OK can we improve the schema?  
Is there a indexing possibility?  
Can we add some stored procedure to help anything things like that, just like prompting yourself with this questions.

 **Speaker 1** 1:40:58  
Umm.  
So Princess suggestion actually, I was also thinking these things I was discussing with Rupali when we were looking at SQL from that perspective.  
But we can have it in a separate discussion rather than that, because this is on the application DB side.  
Most of them, obviously I would want to bring these materialized views and views because that will change the a lot of logic also right on the.  
So this one I would want to keep it and this discussion, but these things, right, what do you think having a separate discussion for these, you know this is application layer we are talking about and this when we are talking about this is we I'm assuming and correct me we be honest to you when Tuhin let me know these things we are talking about the existing application DB and just using it if we can optimize it there itself and these are you know good to have in the back pocket if we go for the approach where.

 **Jubaed Prince** 1:41:28  
Sure.

 **Speaker 1** 1:41:53  
We are having directly from the application DB and I do want to keep this in picture for certain, for example columns or anything that we would think later and we realize no, this is not something we want to have a middle layer and we just want to directly fetch it from in a stored procedure or in a view we want to fetch it from the application DB itself.  
What do you think?  
Am I going in the right direction?

 **Jubaed Prince** 1:42:15  
You are.

 **Tuhin** 1:42:15  
Little bit of a little bit of additional, right.

 **Speaker 1** 1:42:19  
Umm hmm.

 **Tuhin** 1:42:19  
So the materialized views and the views are on the my cycle 8 reporting DB.  
We should come up with a name for those two things.  
Let me just call it on my sequel late, right, because the other one is 5.7, right?

 **Speaker 1** 1:42:28  
Hmm.  
OK.

 **Tuhin** 1:42:31  
So the vision materialized fees are the way we kind of were thinking was on the does it make sense on the my sequel laid the database.

 **Speaker 1** 1:42:33  
Ah.  
Hmm.

 **Tuhin** 1:42:42  
The reason is there is another use case.  
If you remember, we talked about there are some third party partners, right where we would have to create certain tables for them and give them access to those views or materialized views, right?

 **Speaker 1** 1:42:49  
Hmm.  
Hmm.

 **Tuhin** 1:42:55  
They all have a slightly different requirement, so from that point of view we may have to look at 2 linking right, because my SQL light doesn't come with materialized view right?  
We may have to add a plugin or something like that, right?

 **Speaker 1** 1:43:05  
OK.

 **Tuhin** 1:43:07  
So from that point of view, Prince, go ahead.  
Sorry, I carry off.

 **Jubaed Prince** 1:43:10  
Yeah.  
No, that's that's a good point.  
And also on to add on top of that, Priyanka, since you're going through the exercise, why not spend few more minutes and just think it through and then keep it there, don't have to prioritize it.

 **Speaker 1** 1:43:12  
Hmm.  
Umm to.

 **Jubaed Prince** 1:43:22  
But when you are going through it it, since everyone is there, I think it's a good opportunity to kind of figure out what is the best approach, what is the approach you're gonna take now.

 **Speaker 1** 1:43:25  
Hmm.  
Umm.  
No, that makes sense actually.  
And after doing also added you added these points right?  
It makes sense to always keep it there when we are discussing at the back of the mind.  
So we asked this question right, are these are these anything that have to do with this report?  
Anything we have to consider, I think let's keep it there itself and we can.

 **Jubaed Prince** 1:43:49  
Yeah.  
And.

 **Speaker 1** 1:43:52  
But one thing I was thinking, what do you think, Prince?  
This is these are the questions we have to go.  
Like for example right now for the points report, how we ask these questions, we'll there will be a process that for other reports also the same questions have to be always considered.

 **Jubaed Prince** 1:44:05  
Yes, yes, this is like a template of questions that will always ask to each report and to add on top of the material as view topic.

 **Speaker 1** 1:44:09  
Yes, but I'll leave it no it. Hmm.

 **Jubaed Prince** 1:44:14  
One thing you might want to think about is, let's say in a report table.  
Imagine this situation.

 **Speaker 1** 1:44:19  
Umm.

 **Jubaed Prince** 1:44:19  
You have a table that has all the data, but then you have another table that has like a detail level data.

 **Speaker 1** 1:44:23  
Umm.

 **Jubaed Prince** 1:44:27  
So it's a like a level two data.

 **Speaker 1** 1:44:27  
Umm.

 **Jubaed Prince** 1:44:29  
So now you need to do a join and show people.  
Now if you do it through materialized view then this joins can be cached right?

 **Speaker 1** 1:44:34  
Uh will be better, yes.

 **Jubaed Prince** 1:44:36  
So.  
Umm, so that's that's those are the situations where this can really help us optimize.  
So just having those things and asking those question and maybe you do it later, but just making a point that OK in points report we can use materialized views to optimize for this this XYZ things like that.

 **Speaker 1** 1:44:46  
Umm.  
Hmm.  
I think I like that idea actually, to keep it keep a track of we don't need any indexing or indexing.  
You have been done for this materialized view for this scenario on this report has to be created as you said, we don't have to create that right now, but we have to put that there for this reporting.  
We discussed this.  
This is what so that later, even if we come back and we have certain things, so so we don't miss anything.  
I think that's a good idea.  
Let's let's keep it here.  
Let's start one.  
You know, kind of we'll record whenever we have this session and we'll do this for every ohh report.  
Uh, so I think this is all in the same workshop.  
So the to do for this is to set up a workshop on Tuesday morning and start with it.  
I'm assuming because this is going to be like parallel work, we would have to just a heads up.  
What do you think?  
So Tuesday and Wednesday mornings we have two workshop so that we cover everything.

 **Tuhin** 1:45:57  
Sounds good.

 **Speaker 1** 1:45:59  
So I'll set up that, OK.  
Everything else we are going on this one right?

 **Tuhin** 1:46:07  
Yeah, only another thing that I, you know, even if we are not doing it on Tuesday and like, you know, she's calm down a little bit. Priyanka.

 **Speaker 1** 1:46:07  
The changing point.  
Umm hmm.

 **Tuhin** 1:46:17  
So.  
But yeah, so the number 4 is a separate thing that we can start looking into in parallel.

 **Speaker 1** 1:46:26  
Yes, that'll be yes.

 **Tuhin** 1:46:28  
Uh, then.  
Numbers UH-5 I think makes sense and we can work together like you know, this is just to test that CDC and like simulate the load that we are seeing in the current production, right, that's kind of the idea behind there no.

 **Speaker 1** 1:46:39  
Hmm.

 **Tuhin** 1:46:47  
So #6 is something I wanted to ask both Rajesh and to you, right you know because this is coming from when we did the POC evaluation about a year ago at this point, right, there were some recommendation in terms of I think how the angular application was pinging the database or even I think you guys had some recommendation around are we rendering the data using like you know the best tooling in angular or like is there room for opportunities?

 **Speaker 1** 1:46:54  
Umm.  
Yes.  
Hmm.  
Umm.

 **Tuhin** 1:47:24  
Priyanka, what I'm grasping at right?

 **Speaker 1** 1:47:24  
Umm hmm.

 **Tuhin** 1:47:25  
If we can shape up couple of seconds there, that's couple of seconds that we should look into, right?  
So maybe we bring in someone who's an angular expired on our implementation.

 **Speaker 1** 1:47:36  
Hmm.

 **Tuhin** 1:47:38  
Someone who has ideas on your end?

 **Speaker 1** 1:47:39  
Umm.

 **Tuhin** 1:47:41  
Then those two people can work together to see if it's a something that can be optimized fairly easily, or if it's if it's gonna take months, then you know, let's not worry about it.  
But if if there is any, if there is any incremental gain that can be done on the angular angular side, right?

 **Speaker 1** 1:47:52  
Umm.

 **Tuhin** 1:47:59  
So that's that.

 **Speaker 1** 1:47:59  
Yeah.

 **Tuhin** 1:48:00  
That was the idea there.

 **Speaker 1** 1:48:02  
Felt like I I'll.  
I'll get back to you on that one.  
Let me discuss with Anand and Ashoka and this one.

 **Tuhin** 1:48:05  
Yeah.

 **Speaker 1** 1:48:08  
I'll come back.  
Let me put that as to the list for me now to get back to you, because I do think that is a good idea to have both the teams like in other projects also where we are developing the front end right, we are making sure that the data team In Sync with that.

 **Tuhin** 1:48:13  
OK.  
Umm.

 **Speaker 1** 1:48:24  
So that because even if you go for entity frameworks or anything, if you are In Sync with data team and you make sure the performance on the back end modeling is also there's on front end, it does help.  
Actually, the work obviously falls on the angular developers but still it does have OK, OK, yeah, definitely.

 **Tuhin** 1:48:35  
Right.  
Yes, yeah, yeah.

 **Speaker 1** 1:48:45  
I'll come back to you about this one for point the same point report I had a question because we did certain like for certain tables we I mentioned that yesterday to you right we have some uh code ready but I wanted to know the end result that we want.  
So for the points report, are we talking about the 7 to 15 tables that we are fetching the data and then generating because that is what I would think that we would generate the raw data there in the you know the clone copy that Karthik created or it's just on the reporting side, you're thinking, just wanted to clarify.

 **Tuhin** 1:49:23  
Say that one more time I didn't fully understand it.

 **Speaker 1** 1:49:25  
So the 22DB site do D's we have one is the application DB that is the clone I'm referring to right where we are getting the data from and then transforming it and putting in reporting DB so the synthetic data might my understanding would be the synthetic data would be generated for the tables that we are using in application DB.

 **Tuhin** 1:49:26  
Priyanka.  
Yeah, yeah, yeah.  
Umm. Mm-hmm.

 **Jubaed Prince** 1:49:45  
Yes.

 **Speaker 1** 1:49:46  
Right.  
Just like.

 **Jubaed Prince** 1:49:47  
Yes.

 **Speaker 1** 1:49:48  
OK, awesome.

 **Tuhin** 1:49:49  
Yep, Yep.

 **Speaker 1** 1:49:50  
Yes, Sir.  
So that that is clear.  
Uh.  
Another thing, can I make this a separate task?  
Because this is and this will be after this discussion, because this will be like a yeah.

 **Tuhin** 1:49:57  
Please go ahead.  
Yeah, yeah, yeah.  
Whatever makes sense, right? Yeah.

 **Speaker 1** 1:50:02  
So I'll have somebody work on this one.  
I have to figure out who will work with this.

 **Tuhin** 1:50:07  
Yeah.

 **Speaker 1** 1:50:07  
OK.  
Yeah, like this this function.

 **Jubaed Prince** 1:50:09  
Up.

 **Speaker 1** 1:50:10  
So now we are just looking at this approach right?  
No JS we can do, but I think that'll be a whole can of worms for one week.

 **Jubaed Prince** 1:50:17  
That we we don't have to do all of them.  
We just have to do one of them or in outside these whatever works visually at this point like for next week.

 **Speaker 1** 1:50:22  
OK.

 **Jubaed Prince** 1:50:27  
It's just something that works that makes sense, but we'll do decide it be a discussion.

 **Speaker 1** 1:50:31  
Umm.

 **Jubaed Prince** 1:50:33  
I I'm sure you you need to you need time to think about this.  
So this is just a like a probable solution, not a uh.

 **Speaker 1** 1:50:38  
Yeah, because.  
Umm.

 **Jubaed Prince** 1:50:42  
So you you just pick one or and then you go with that, but you you discuss that why you are choosing that.

 **Speaker 1** 1:50:45  
Hmm.  
Hmm.  
So, so, OK, that that makes sense, right?  
At least have a bare bone one.  
So I'll come back to this.  
I wanted the 4th one right, so I didn't mention here.  
I want, I think what we discussed and all of us almost are in agreement that we should configure this debezium, you know, in a data sync mechanism that syncs into the S3 bucket.

 **Jubaed Prince** 1:51:10  
SO55 is the part of four, by the way, the five is that data sync mechanism for four.

 **Speaker 1** 1:51:15  
Yes, it it, yeah.  
So no, actually that is what I'm confirming.  
Right Lambda function.

 **Jubaed Prince** 1:51:22  
Yeah, we'll do that Data thing.

 **Speaker 1** 1:51:24  
You you you can actually.  
Yeah, so Data sync?  
Uh, can actually, we were discussing earlier, right prints that we can have the configuration to sync the data into S3 and the Lambda function can be triggered on that.

 **Jubaed Prince** 1:51:34  
Yes.

 **Speaker 1** 1:51:38  
You know, getting the data there and it will just quickly transform and put it into reporting DB, right whenever there is a change in file.

 **Jubaed Prince** 1:51:39  
Sure.  
Sure.

 **Speaker 1** 1:51:45  
So that's why I'm putting it as separate this one.  
I'm adding it to this one so that rather than having reporting DB as the our aim of sync that will be the S3 reporting S3 we can call it right and this function will pick what I'm thinking is from the S32IN automated way.

 **Jubaed Prince** 1:52:03  
Report, baby.

 **Speaker 1** 1:52:03  
Can we?  
Yeah, enter the reporting DB.  
Am I?  
Is that something that you so your boarding S3 bucket put it so that it's clarified, yes.

 **Jubaed Prince** 1:52:07  
Yes you are.  
You you are right, yes.

 **Karthik Raghu** 1:52:13  
I have one question, one who suddenly straight.  
So because we are see our initial thought process was just to there was just one hop right like one hop in the sense between the Kaka and the my sequel 8, there's one hop, right?

 **Speaker 1** 1:52:17  
Umm.  
Hmm.  
Hmm.

 **Karthik Raghu** 1:52:35  
But right now I think we are talking about 2 hops.  
Uh, which is 1/2 is at the three bucket level and then the other half is at the SQL level, right?

 **Speaker 1** 1:52:39  
Yes.  
Yes.

 **Karthik Raghu** 1:52:47  
So there are two things that are that we are doing in order to push the data into my sequel, right?

 **Speaker 1** 1:52:48  
Correct.  
Umm.

 **Karthik Raghu** 1:52:55  
So in a real time like you know if you if if at all you want to capture CDC how long does it take like because we are like you know it has to initially get the change data that is processed one and then that change data has to be transformed into a Jason and then pushed into S3 bucket that is 2 and then there has to be some other Lambda function or maybe an Azure function which will kind of continuously monitor the bucket or the storage and then it will kind of push all the data into.

 **Speaker 1** 1:52:56  
Yes.  
Umm.  
Hmm.  
Hmm hmm.  
Umm.  
Hmm.  
Umm.

 **Karthik Raghu** 1:53:30  
All the and then convert or transform the data and then push it to my sequel.  
So how long do you think all this process will take?

 **Speaker 1** 1:53:36  
Umm.

 **Karthik Raghu** 1:53:38  
Let's say there is one change.  
Uh, but in production, obviously, I think there will be multiple changes that happen every second.

 **Speaker 1** 1:53:45  
Hmm.

 **Karthik Raghu** 1:53:46  
I think we are getting somewhere around the 1000 requests on our USB which is like, you know, it's pure of them could be get.

 **Speaker 1** 1:53:50  
Umm.

 **Karthik Raghu** 1:53:54  
But still I'm just trying to give you an idea, right.  
So the data will continuously.

 **Speaker 1** 1:53:57  
Umm.

 **Karthik Raghu** 1:53:59  
What you say, stream?  
Yeah, as it in production.  
So how long do you think it will be this will take?

 **Speaker 1** 1:54:05  
Ohh so actually that is good.  
Ohh, I also thought about something on that.  
Ashok, you would be a better person to answer, at least from you know.  
Consuming CDC from my SQL to S3.  
How much time do you think that'll take for a certain amount of maybe do you have it an idea?  
Ashok.  
I'll get back to you on that, Karthik.  
Right.  
I think Ashok has dropped off for another call.  
I will get that answer for you.

 **Karthik Raghu** 1:54:44  
Yeah.

 **Speaker 1** 1:54:45  
And because you mentioned that I've.

 **Karthik Raghu** 1:54:46  
Because we discussed about a lot.  
Yeah, we discussed about Lambda is event based right?  
So in other words, like basically, I was thinking that you don't need an event like it has to be continuously running because I 100% sure that every second there will be thousands of transactions coming in.

 **Speaker 1** 1:54:52  
Yes.  
Umm.  
Umm hmm.

 **Karthik Raghu** 1:55:04  
So that is where we need to think.

 **Speaker 1** 1:55:04  
Umm.

 **Karthik Raghu** 1:55:07  
Think.  
Think it through.

 **Speaker 1** 1:55:07  
No, actually that's what I'm thinking.  
That is the one thing that I I also realize when you were talking right, if you we were taking it into the Lambda function, we were triggering a Lambda function and it was taking that directly.  
That would be the trigger.  
If we go trigger event like as you said, every second there'll be new change.  
I'll have to look at that right in the ESG work at how generally Lambda is triggered as event based right event will just trigger the Lambda and we did talk about parallel processing and right there's a limitation how much parallel lambdas can work, although it will so and how much time actually is your for Lambda is taking because it's only table so it should not.  
It should take a few seconds only to update, but still we have to consider that so that it does not cause like overlapping and does not have a problem with the data.

 **Karthik Raghu** 1:55:54  
Yeah.

 **Speaker 1** 1:55:54  
Do you think that'll be a good discussion to have before we start doing that and putting it in, yes.

 **Karthik Raghu** 1:55:58  
Yeah.  
Keep this in mind because this is this is like was thinking too much and then maybe I'm thinking too much as like but.

 **Speaker 1** 1:56:04  
No, no, no.  
You you actually, because I remembered yesterday.  
Yesterday, when I the presentation that I was, we were going to write and I had this discussion with Ashok and because we were picking up from the file for the Lambda, we were like, OK, 5 minutes.  
Let's not the Lambda as well as the SQL also right.  
If you remember, I said near real time for 5 minutes because it's like what do you mentioned every second there's a change.  
Do we want to run those things every second, or do we just automate it using Kafka?  
Debian.  
So let me have this discussion with Ashok and come back for Tuesday with an answer, right.

 **Karthik Raghu** 1:56:41  
OK.

 **Speaker 1** 1:56:41  
Uh, so I'll I'll add this here as a comment for myself. Umm.

 **Jubaed Prince** 1:56:49  
Umm I want to add something here.  
Also, as Karthick mentioned, it adds one hop and the cost of 1 hop is giving us a way to capture the data and batch process them so that fail.

 **Speaker 1** 1:56:52  
Umm.  
Umm.  
Umm.

 **Jubaed Prince** 1:57:06  
There is some sort of a fail safe mechanism there.  
That's why I Ashok kind of was proposing that now The thing is, if if Ashok can solve that problem without having another hop, that would be interesting.

 **Speaker 1** 1:57:10  
Yes and yes.

 **Karthik Raghu** 1:57:12  
Correct.

 **Jubaed Prince** 1:57:19  
And I believe it is probably possible.  
I don't know.  
Theoretically speaking, maybe Bin Log has some time stamp and stuff and we can just go back and see OK, what was our last sync and go back and match that?  
I'm not sure how that gonna work.  
Ashok will know it better, but I think it's it's a very important discussion to have that.

 **Speaker 1** 1:57:35  
Hmm.

 **Jubaed Prince** 1:57:41  
Should we add these S3 cause it's gonna cost us.  
Money.  
It's gonna take more time, but it's gonna give us some sort of stability.

 **Speaker 1** 1:57:46  
Umm.

 **Jubaed Prince** 1:57:50  
Now how can we so so that trade off decision?  
I think we we need to do it all together, but before that you you guys should like kind of run some experiments and play around and come back.

 **Speaker 1** 1:57:57  
Yes.  
Actually brings.  
That is, that is why I'm saying that I will come back to you because I was also at the same thought process and that's why earlier also we talked about like 5 minute delays.  
Is it OK?  
Right.  
Is that the trade off?  
We are OK with the right because in 5 minutes we can pick up everything and then the Lambda function can do the delta right?  
But I want to make sure that if we are then at the approach to trigger Lambda will be different right?  
And other thing I want to make sure that even if we skip that that we say no that we don't want that trade off, we want directly real time reporting DB update and then is it possible using the Lambda with the transformations right?  
Because syncing it directly into the DB, yes that is there, but transformations I'll have to confirm with the I'll have a brainstorming with Ashok on that or come back to you.  
Right.  
So all these things I think now I have all of them, at least I should have, how it'll do and we can decide on the trade off.  
So do you think umm?

 **Tuhin** 1:58:58  
Yeah.  
Only thing Priyanka, just as a as a guiding principle, right?

 **Speaker 1** 1:59:01  
Umm.

 **Tuhin** 1:59:03  
I don't want to move into batch process unless we absolutely have to and exhausted all other option.

 **Speaker 1** 1:59:07  
Yes.

 **Tuhin** 1:59:08  
Yes, that's that's just one to make sure.

 **Speaker 1** 1:59:09  
No.  
And yeah, no, not the.

 **Tuhin** 1:59:11  
Yeah, yeah.

 **Jubaed Prince** 1:59:12  
Yeah, that's I am also with Tuhin on this, if we if we need more time to figure that out, I think that's a good investment of time.

 **Speaker 1** 1:59:15  
Yes.  
No, I actually agree to that because I think you're requirement does not just few of the reporting which does not make sense to go for batch processing.

 **Tuhin** 1:59:21  
Uh.

 **Speaker 1** 1:59:28  
It should all be real time only like and we should try to get as much CC as possible, right?  
Maybe what Tuhin earlier in today's call I mentioned like 5 minute was like OK every 5 minutes you have a, you know what is running there won't be overlaps and there won't be.

 **Tuhin** 1:59:41  
Yeah.

 **Speaker 1** 1:59:46  
Because I'm also thinking about data quality right that the reports don't override each other, we have a fail safe, but if the something is writing the same thing, it should not let it write in your reporting DV.

 **Tuhin** 1:59:49  
Yep.

 **Speaker 1** 1:59:57  
That kind of thing that is going in and because that you are already facing that problem, we don't want to replicate that.

 **Tuhin** 2:00:05  
Yeah.

 **Karthik Raghu** 2:00:05  
Yes, these are problems basically right, like the having having a fail safe.

 **Tuhin** 2:00:05  
Yeah. Make.

 **Speaker 1** 2:00:08  
Umm.

 **Karthik Raghu** 2:00:09  
We need to cover that.  
Like, you know, having real time CDC as much as possible, we need to cover that, making sure that the solution works on AWS and Azure.

 **Speaker 1** 2:00:11  
Yes.  
Hmm.

 **Karthik Raghu** 2:00:18  
That is one more thing that we need to consider.  
So all these are considerations.  
So yeah, wanted to put in.  
That's like and then the solution should actually take all these boxes.  
Uh, you know that that's what we call it as a solution, right?  
So yeah, you may have to think like note down all the on the things on the, on the checkboxes that we need to the you know pick and then come back with solution which covers all these things.

 **Speaker 1** 2:00:37  
Umm.  
Yeah.

 **Jubaed Prince** 2:00:50  
Sachin had something to say.

 **Speaker 1** 2:00:50  
OK.

 **Sachin Naikwadi** 2:00:51  
Yeah, just one question on the failure of debasement series.  
So data will not go to Kafka right then, but our real application have that data.

 **Speaker 1** 2:00:57  
Umm.

 **Sachin Naikwadi** 2:01:02  
So how to handle that singing the lagging of that data or the failure type?

 **Speaker 1** 2:01:04  
Umm.

 **Karthik Raghu** 2:01:13  
Yeah, I think that's. Yeah.

 **Speaker 1** 2:01:13  
Can you repeat that, Sachin, if?

 **Anand sivaraman** 2:01:16  
So we we need to, we need to employ a retry mechanism that we need to employ a retry mechanism.

 **Karthik Raghu** 2:01:16  
Yeah, it's the same question, right?

 **Anand sivaraman** 2:01:22  
Kafka has the retry mechanism capability, but how many times do we want to retry?

 **Speaker 1** 2:01:23  
Hmm.

 **Anand sivaraman** 2:01:28  
Is a business decision that we have to take right and after two or three times retry it doesn't happen, then we'll have to push it to an exception file.  
Ideally have a user have a team review that why the data that particular update did not come through and then do a push it through a batch update.  
So this is a common process that but gets followed there, but we'll have to discuss based on a case to case basis.

 **Speaker 1** 2:02:00  
OK.  
So I think the these are all the points we'll come back and this is another thing that we can discuss a.  
OK so.  
Are we good with the museum connection like all the all the considerations that we have to take and we have to come back and then finalize the approach on that using the Lambda function or if something else that can transform directly from Kafka if we need that but we do want to also discuss when we come with answers about the feel safe right on both sides on data quality as well and the any failures on Kafka side and putting everything on S3 just having a always having a current file or current data.

 **Jubaed Prince** 2:02:51  
Yep.

 **Speaker 1** 2:02:51  
Yeah.

 **Jubaed Prince** 2:02:52  
I also want to yeah.

 **Speaker 1** 2:02:52  
And we have already, yeah.

 **Jubaed Prince** 2:02:55  
Like I I since we're running out of time, let's also discuss a little bit about 8:00 and 9:00, just to surface that.

 **Speaker 1** 2:02:58  
Hmm.  
Sure.  
Yeah, I think 6 and seven.  
We have already discussed and let's go to eat every anything else, anybody has still .6 for now.  
OK.  
Let's move on to 8, which is like good to have till 29th, right?  
But yeah, Prince, you wanna take this one?  
State an expert excel requirement, OK.

 **Jubaed Prince** 2:03:31  
I'm so far Excel, I think to him, if you can tell a little bit that would be.

 **Tuhin** 2:03:35  
Yeah.  
So yeah, so right now we're running Cron jobs.  
Essentially, when the in current situation that's emailing the reports out to the end user and the Crown job right now, you know as you can think about when the application is hitting the query, the same thing is happening on the Cron job as well, right?

 **Speaker 1** 2:03:50  
Umm.

 **Tuhin** 2:04:00  
And we have a real issue right now in terms of consistent SLA in terms of how much time it takes our system to generate the report.  
So the goal here is that OK, we are solving the report UI problem right and the data is getting getting denormalized, getting rolled up as much as possible.  
We are putting it in a uh, you know, we are looking at views and etcetera, what have you right.  
We have done all of this optimization for the UI and at the end of the day right for the end user reporting is also an important factor.  
Like the report export that our end user gets.  
So my thought process is that is that any tooling that or anything like that that we should be looking into which lets us like, you know, kind of move away from the uh background Cron jobs that are running to send those reports out and make make make the processing a little bit more predictable, right.

 **Speaker 1** 2:04:57  
Umm.

 **Tuhin** 2:05:03  
And that's that's kind of the question, right?  
I don't know if there are.  
I'm I'm sure there is right there could be you know, you know plugins or opensource what have you or maybe something we have to write right?  
Is there something that lets us export?  
Uh reports using the optimized database that we are creating for reporting and you know we should, we should think about it because that's the other part of the user experience that we shouldn't.  
We shouldn't ignore and there are some safeguards and stuff like that.  
We can talk about right now.  
There is no limit in terms of how much data you can export.  
The people can try to export 100 million record and our current job will try to try to export that record, right.  
We are open to looking into like, you know limiting, like saying like we're not letting users export over certain amount of records or what have you right.  
You can put a hard limit at like a 1,000,000 record or something like that, or 500,000 records.  
I don't know what that limit is, but essentially a tooling exploration saying like OK, is there a better way to handle exports?  
So that's kind of the problem statement there.

 **Anand sivaraman** 2:06:18  
Yeah.  
Yeah, Q and I think, I mean, I'm not trying to solve the problem here, but that can be that can be easily achieved in two ways.

 **Tuhin** 2:06:23  
Yeah.

 **Anand sivaraman** 2:06:26  
One is, you know, pulling the the end report that is, that is being aggregated, pushing that and storing it as a parquet file in ADLS Park is a very low compressed, high compressed file with low cost.

 **Tuhin** 2:06:33  
Umm.

 **Anand sivaraman** 2:06:41  
So the in in in Azure Data Lake services are.

 **Tuhin** 2:06:42  
Umm.

 **Anand sivaraman** 2:06:45  
Yeah, or in BLOB storage even or in S3 and then directly connecting that to directly connecting that to a service like Power BI or even AWS quick site, right?

 **Tuhin** 2:06:49  
Umm.

 **Anand sivaraman** 2:06:56  
Where the you know the we can you can directly connect that and allow the user to self serve those 1,000,000 records within Power BI.

 **Tuhin** 2:06:56  
Umm.  
Yep.

 **Anand sivaraman** 2:07:04  
Not a visualization, but purely as a table, right?  
With all the pagination and whatever right that can be possible given.  
I don't think that's a huge challenge, but the only only thing is how do we make sure that whatever is currently happening in the UI in the angular framework does not conflict with that and allow the user to probably use this power BI console or a BI tool console for that is something that we can discuss.

 **Tuhin** 2:07:13  
Yeah.  
Umm.  
Yeah, only one one correction there.  
And then this that we're talking about, not necessarily UI interface but a Download file which could be a CSV, could be an excel that's that's all.

 **Anand sivaraman** 2:07:40  
Yes, yes, sure, sure. But.

 **Tuhin** 2:07:42  
That's all right.  
So it doesn't have to anything happen.  
Anything on the UI but it just a physical file that gets downloaded.

 **Anand sivaraman** 2:07:47  
And that that's correct.  
So that's even probably better than downloading that into.  
Umm, some kind of a file and then exporting that into a CSV?  
Shouldn't be a problem too.

 **Tuhin** 2:07:59  
Yeah.  
Perfect.  
Thank you. Yep.  
And the last one, automation testing, essentially one of the perils that we ran into that Databricks project was that, you know, not only we have to, how do I say this not only we have to look at the data frame dimension, but we do need to create for the export file and on the UI I don't know how easy it is and internally maybe we'll look at it if you have guidance you let us know as well.  
We need to put some automation testing there because we do have to absolutely have to make sure when we go release these things in production, especially for the high touch customers who have complained about databricks, we don't run into this issue and manually doing that for millions and millions of records because it may not be a feasible thing to do.  
So I'm just kind of thinking about like, you know for the file export, maybe it's easy, maybe you write a Python script, whatever that takes an export of the existing report and the new report export that we just kind of talked about and it just does a comparison, we get a piece of mind there, but we need to speed up this testing process.  
Yeah, the UI testing maybe can be done manually, but at least for the report exports that we need to do, we may have to write some sort of an automation script that we can test.  
Yeah.  
A large number of customer sites.  
Uh to make sure that we're not seeing data quality issue and things of that nature.  
So that was kind of the motivation behind it.  
There may be other testing, best approaches, best practices, Prince.  
You may want to add anything, but that was the that was the last point there, yeah.

 **Anand sivaraman** 2:09:52  
Yeah, got it.

 **Jubaed Prince** 2:09:52  
Yeah, so that I.

 **Anand sivaraman** 2:09:53  
Got it.  
Sorry, go ahead.

 **Jubaed Prince** 2:09:55  
I will also add couple of lines here that the goal here is to have parity between the current setup and the new that we are doing and ideally we think that being able to create some scripts that that does that for us saves a lot of manual time.  
And I think when it comes to testing, this should be the first thing we should be thinking about.  
How can we automatically test this and then we will do manual testing but but the automation tests should be the primary thing that we should be thinking about when it comes to any of this report.

 **Anand sivaraman** 2:10:36  
Yeah, got it.  
So so we can we can share some thoughts around it.  
We work with tools like Great Expectations, dot IO or DQ or Apache Grifin which act as both data quality checking tools, which allows you to inherently create rules when integrated with them and then help in quality testing of data.  
But we'll also, I understand from queuing that it's not just the quality, it is also the volume.  
So we'll we'll probably have to add some additional scripts, but these three tools are something that we have seen work well as well.  
So we can we can also look at that.  
Otherwise, there are dimer doesn't pipi packages that are available for testing data, right?  
So that is something that we can explore as well, yeah.

 **Tuhin** 2:11:21  
Yep.  
Yeah.  
Thank you.  
Yep, makes sense.  
Sounds good.  
I know we are overtime, just want to see if there is any final thoughts from anyone.  
Before we break.

 **Anand sivaraman** 2:11:42  
We're good tune.

 **Tuhin** 2:11:46  
I just wanted to thank everyone and I think Maddie are here.  
Please feel free to add your thoughts as well.  
Cannot say how appreciative like you know, cannot overstate it of everyone's effort, both internally and Annex Cloud, as well as & on the and the KANINI team for pivoting.  
And I think this session gives me personally a lot of confidence that we are gonna be successful with this approach and you know, and I think once we figure out the plan, it's just a matter of execution.  
I'm sure they'll go as smooth as it can and thank you so much for, you know, pivoting and meeting us for our needs are and you know, we really really appreciative to everyone internally who have already done the work.  
And obviously KANINI team for pivoting.  
Thank you.  
Again, thank you so much.

 **Anand sivaraman** 2:12:46  
Thanks a lot, tune and Team.  
Excited to work on this so good problem to solve.  
Hope to work with you all and get this and get this done so.

 **Tuhin** 2:12:55  
Yeah.  
Alright, thank you guys.  
Have a good weekend.  
If I don't talk to you and Priyanka, I think we can cancel the sync that we have in the morning.  
So then we'll just catch up and happy Holi to everyone.

 **Speaker 1** 2:13:04  
Yeah.  
Thank you so much.  
I hope everyone has a nice weekend as well.

 **Rajesh Bute** 2:13:09  
Thank you doing.

 **Jubaed Prince** 2:13:12  
Thank you all, bye.

 **Tuhin** 2:13:12  
Thank you. Bye.

 **Sachin Naikwadi** 2:13:14  
Thank you. Yeah.

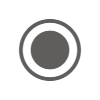
 **Rupali Ghosh** 2:13:14  
Thank you all.

 **Anand sivaraman** 2:13:14  
Thank you.

 **Muniaraj Thangavelu** 2:13:16  
So.

 **Speaker 1** 2:13:16  
Thank you.

 **Rajesh Bute** 2:13:24  
Thank you.

 **Priyanka Kochhar** stopped transcription