**AnnexCloud-Kanini Connect-20240315\_182724-Meeting Recording**

March 15, 2024, 1:00PM

47m 9s

 **Tuhin** 0:03  
Yeah, right.  
Nice to meet you.  
You know, I will.  
We'll do a quick round of intro from our end.  
Uh, like you know, Kartik, do you wanna go first or I can? Yeah.

 **Karthik Raghu** 0:17  
2IN what was that?

 **Tuhin** 0:19  
No, like you know, we're just like quickly doing an A quick intro to naraj and whose has joined our team. Yeah.

 **Karthik Raghu** 0:24  
Yeah, sure, sure.  
Sure.  
I high nitrogen so myself, Karthik, so I'm from.

 **Natarajan Sundararajan** 0:28  
Like.

 **Karthik Raghu** 0:32  
I'm a DevOps manager, so I'll be your point of contact for with regards to any kind of access related stuff or any kind of, you know, uh information with regards to how the environments.

 **Natarajan Sundararajan** 0:41  
Yeah.

 **Karthik Raghu** 0:44  
Uh, spin.  
Spin like spinning like you know what kind of databases you have?  
Connectivity, you know, right, listen to me for any kind of recommendations, right.

 **Tuhin** 0:51  
Mm-hmm.

 **Natarajan Sundararajan** 0:52  
God. OK.  
Sure, sure.

 **Karthik Raghu** 1:00  
True.  
Yeah.  
So I'm based out of Michigan, by the way.  
I think so.

 **Natarajan Sundararajan** 1:03  
OK.

 **Karthik Raghu** 1:04  
So Carl is from central time, I think is from Texas.  
So so I think I'll.

 **Carl Droste** 1:09  
Yeah.

 **Natarajan Sundararajan** 1:09  
Yeah, yeah.

 **Carl Droste** 1:12  
Dallas Richardson.

 **Tuhin** 1:17  
And Carl, do you wanna just quickly introduce yourself to another Rajan?  
He joined our team the in this engagement recently.

 **Carl Droste** 1:30  
Uh, and what else?

 **Tuhin** 1:33  
Uh Natarajan is a new team member from.

 **Karthik Raghu** 1:35  
Your your voice is kind of muddy to him.  
Maybe like you know, it's just that's it.

 **Tuhin** 1:38  
Ohh OK.

 **Carl Droste** 1:39  
No, no, it's alright.  
I I just.  
OK, so now Rajan is is.

 **Speaker 1** 1:45  
So call Natrajan is based out of US and he has just joined committee team on the management side.

 **Tuhin** 1:46  
Yeah, what?

 **Carl Droste** 1:46  
But.

 **Natarajan Sundararajan** 1:46  
Garland was here.

 **Carl Droste** 1:52  
OK.

 **Speaker 1** 1:52  
And so just wanted to introduce you to all of you.

 **Carl Droste** 1:55  
OK.  
All right, yeah.

 **Speaker 1** 1:56  
You'll be seeing him a lot in your meetings.

 **Carl Droste** 1:59  
OK, OK.  
Yeah, I I'm I I'm.  
But I've been with Annicka from last couple years now and I I serve in the capacity in the US side, you know, engineering management.  
Ah, director of engineering here in the US side of things.  
So I try to help in between.

 **Natarajan Sundararajan** 2:19  
OK.

 **Carl Droste** 2:24  
It connects the the dev side to the US CIKM CSM and all of the things over here and make sure that we're we're we're In Sync.

 **Natarajan Sundararajan** 2:38  
Yeah.  
OK, cool.  
Yeah.

 **Tuhin** 2:44  
You and yeah.

 **Natarajan Sundararajan** 2:44  
Thank you.

 **Carl Droste** 2:45  
Nice to meet you.

 **Natarajan Sundararajan** 2:46  
Yeah, I seem to meet you.  
Nice to meet you.

 **Tuhin** 2:49  
Can you guys hear me alright right now?

 **Speaker 1** 2:52  
Yes then.

 **Tuhin** 2:53  
OK.  
Yeah, I had to turn off my Wi-Fi and the the mobile data works better.  
Anyhow, I I naranjan my name is Juan like I'm I'm based in Eastern Time.

 **Natarajan Sundararajan** 3:02  
Yeah, right.  
Umm.

 **Tuhin** 3:07  
I'm the head of product here at the next cloud, essentially looking at looking after the core product offering as well as integration.

 **Natarajan Sundararajan** 3:15  
Umm.

 **Tuhin** 3:15  
And obviously I interface with Carl and you know, Kartik and rest of the engineering leadership team making sure we're working in the right priority, innovating, et cetera.  
And you know, kind of been championing the community engagement internally on our end.  
So that's kind of my role here, yeah.

 **Natarajan Sundararajan** 3:35  
OK.  
Thank you.

 **Tuhin** 3:40  
Before I get started and I did receive both of your email and we can talk about that.  
Sorry, I was just crazy busy yesterday, so I didn't respond to you, so apologies for that.

 **Anand sivaraman** 3:52  
No, no problem at all to you.  
And I I wasn't sure if my mail reached out reached you because we have.

 **Tuhin** 3:54  
Yeah.

 **Anand sivaraman** 3:58  
We've been revamping our outlook and Microsoft installations here, so just wanted to be sure.

 **Tuhin** 3:58  
Yeah, yeah.  
Like, yes. Yeah.  
Can't go.  
We can get started.  
Sorry for the sorry for that.

 **Speaker 1** 4:12  
No issues at all.

 **Tuhin** 4:13  
Yep.

 **Speaker 1** 4:14  
So I have two presentations today.  
One is the one is part.  
We have already covered on Wednesday, right?  
So let me start the first one.  
I'll just go through the WSR which has the deliverable as well on so.  
Let me quickly go over to the Blue star for this week.  
So we have the same agenda.  
This is the same as Monday board, so this week this was from the past week.  
This week we had five agendas that we decided starting right.  
That was so the cost analysis which we went through and the file is attached on the in the findings section in the in the in the WR as well as in Monday board.  
Same goes for the scrub data migration approach.  
I think the main call you might have already received that from Anand yesterday, but we have also added to that same slide and I'll go over it after this WSR.  
Uh.  
Then we had the approach we had to provide an approach to deal with the duplication and missing data, but we do need to, you know, I think we went over it last week also and this week also that we have certain points we can now we can put them in.  
I have put couple of them in the slide but it seems like the duplicates and missing data are being introduced into the bronze Liv and the data is being taken and this needs to be, you know addressed as a approach how we are going to actually code whether we are going to just fix small fixes in the code or we're going to, you know, just take a new approach and that is the same thing that when we talked about Wednesday that we'll test out on a small code on production to see if data entire raw or.  
Changing the current one, which will take less time and will approximate cost because the cost is not that won't be as much volume as it is currently right in production, but approximately we can give you an approach which will be better, right?  
So this is this I'm looking on this.  
We'll test out next week about the increasing the concurrency.

 **Tuhin** 6:10  
Mm-hmm.

 **Speaker 1** 6:14  
This has been parked because till we have data volume to actually test it effectively, whether it's reading or whether it's, you know, alongside with the conferencing settings, how we can turn code itself, right?  
This will be a change in code, so this is something that can be again done in in the bronze layer and then going forward in the silver layer.  
But this does need data volume to test any efficiency smaller approach we're always tell how to do the training, so that's why we have parked this.  
What we have also attached the storage cost right region wise.  
How how much the database and this all analysis time we have very focused on data bricks, right?  
And based on the resource groups for data bricks so that we can see how much changes or what is currently costing you or the performance on so it's just costing, right, how much it's costing on you know, on adls and DBFS.  
So that is there, these are the findings from the last week.  
They are same as the Monday board, the new ones for this week.  
I've been putting this to PD and I'll go through them.  
So I'll go through the VPN itself, right?  
So this was the one of the decks that we shared.  
Now this deck has first we we have a summary of what we saw this week and what we analyzed right up like a shot some and checks and sums right.  
And also the more detailed cost analysis with exact numbers and we have changed a week.  
So one week is from 7th when we actually did the changes half after half day and more changes when they were done with, you know, on a day before by offshore team as well.  
So this is this has all the cost for that right till 13th because we wanted to have a fair day to day and we have put the changes and everything that we already covered job scheduling also this has the same data obfuscation approach.  
The three approaches that we suggested in this slide is attached in the WSR itself and data quality.  
As I said, we have put couple of them for you to see where the issues are in the code and what how it's actually duplicate time being created.  
The data quality issue where it's coming and a single page storage content analysis which we have to go forward and we will actually put more analysis after talking to the offshore team how they are storing the data and how is it impacting and how it can be changed, right.  
So just a quick summary again, we I already mentioned that we compared the seven day period since March 7th when you know the changes were made on the POC cluster, we did the all purpose job, compute cluster changes and later on 13th even the offshore team changed the uh the SQL warehouse cluster.  
One of them was reduced right as the suggestion, so we saw $160 to $500.00 per day cost savings, right.  
Per day cost savings, right.  
And total for that this week from 7th to 13th we will see a cost savings of $2000, right?  
Just in POC itself, on production we just changed the schedules of the jobs and so that they don't face any issues, right.  
They don't face any failures and still have them optimum one and this was the savings for each of them.  
So HFC, we saw from five 1000, I think I, sorry, I'll change this number, it is in there for the US it's the other one.  
I'll.  
I'll go forward and and then you was the only reason that we did not say see much saving it might be because I think the offshore team was working on the Harrods, you know, requirements and the run times were increasing and they needed another run to run.  
Harrods continuously.  
So, uh, so when we I-1 disclaimer, I do not have the screenshot from 7th of March, but I do have from a 13th so show the comparison.

 **Tuhin** 9:54  
OK.

 **Speaker 1** 10:04  
But we did see the cost forecast from that.  
We remember from March 7th that was 82K for March.  
Change to 67.9 today $1000 OK uh that we have put the three data frustration approaches and again this is the same thing, right we identify two causes for duplicates and storage uses it is for bricks cup databricks covered, right.

 **Tuhin** 10:13  
OK.

 **Speaker 1** 10:28  
So let's quickly go through the POC until the changes are there.  
We have gone through them, right?

 **Anand sivaraman** 10:32  
Hip hip, Priyanka, Priyanka just just one second.  
I think I think we covered a lot of stuff, so let's just hear back from Carl and you in and Karthik, if there are any questions, first please.

 **Speaker 1** 10:41  
Sure.  
Good point.

 **Tuhin** 10:47  
So far I'm good.

 **Karthik Raghu** 10:48  
Yeah, I was.  
Uh, yeah, thanks for thanks for pointing it out.

 **Tuhin** 10:49  
The fact that the.

 **Karthik Raghu** 10:53  
Yeah.  
So I was looking at the done and then in the parking lot section, right.  
So with regards to scrubbing approach so.

 **Speaker 1** 10:57  
Umm.

 **Karthik Raghu** 11:02  
Should.  
Did didn't we actually discuss to, you know include that?  
Scrubbing approach as part of this exercise or we still have to work on.

 **Speaker 1** 11:15  
No.  
So the what we included in this Karthik was the approaches unless As for what kind of approaches we can take for you know for scrubbing the scrubbing is a different separate separate exercise in itself.  
And for that, we will have to wait for him to come back.  
What approach ohh you want to take right tohen and you and definitely and then we can, you know work on that.

 **Tuhin** 11:38  
Yep.

 **Speaker 1** 11:41  
So I mean maybe you can add something on that one.

 **Tuhin** 11:44  
Yeah, I think, uh, like, I think the yesterday we kind of talked about like, you know, not yesterday, the day before I think the and Priyanka, correct me if I'm wrong, the idea was to understand the pros and cons of both approach and like you know from a timing perspective more so than the cost that how much time you was gonna add etcetera.  
And I think Anand had sent out some suggestions yesterday.  
I haven't fully had a chance to like go through it and Priyanka, you did mention that we're going to talk about that at the end on the second part.

 **Speaker 1** 12:19  
Yes.

 **Tuhin** 12:20  
So that's where we are, Kartik.  
I don't think we have finalized it, yeah.

 **Karthik Raghu** 12:23  
OK. OK.  
Yeah.  
Thank you.

 **Anand sivaraman** 12:27  
Yeah.  
And and and two in for like you rightly mentioned.  
We will walk through each of those points in the email along with the attachments as well.

 **Tuhin** 12:35  
Yep.

 **Anand sivaraman** 12:35  
So we had given three different approaches with scrubbing and without scrubbing, and the pros and cons for us to discuss, evaluate, and then solicit your approval.

 **Tuhin** 12:42  
Yep.

 **Anand sivaraman** 12:47  
And you know, before we move forward, Sir.

 **Tuhin** 12:49  
Yep, Yep, sounds good.  
Thank you, Sir.

 **Karthik Raghu** 12:52  
Uh, you know, with regards to production, so I went, when do you plan to optimize production or do we even have have that in the plan because I think especially US, I think it's been a major contributor to the cost right, like the USB cluster.

 **Speaker 1** 13:08  
Yes.

 **Karthik Raghu** 13:09  
So I think, do we have anything this week or what is the plan for that?

 **Speaker 1** 13:17  
So, Kartik, we were thinking about taking that next week, right?  
Uh, yeah, but depending on how we go for the, you know, the bronze layer approach and also we have definitely next week we'll start looking at the numbers of the you know of homotypic or how much memory you'd realization you see continuously what is the window we can keep.  
But so that we can optimize admins as much as we can, right?  
I get where you're concerned.  
Is, so I think we can take up for the next week, at least for the configuration side.  
Do you think that is a good idea?

 **Karthik Raghu** 13:53  
Yeah.  
Yeah, yeah, yeah.  
At least we can reduce it, right?  
Like we know, what is the current state right now.  
So instead, if even like before even fixing the current state, let's try to optimize it.

 **Speaker 1** 13:58  
Yes.  
Hmm.

 **Karthik Raghu** 14:04  
I don't think it takes much time for you to analyze, right?

 **Speaker 1** 14:06  
No, definitely so.

 **Karthik Raghu** 14:06  
Like you just look up to just provide your analysis and then you know the changes.

 **Speaker 1** 14:10  
So.  
So I hope you'll see.  
Actually I went through all those numbers to suggest any numbers for scheduling I could do right?  
It'll take me.  
Give me one or two days to antice to exactly so because I don't want those two queue up the jobs to queue up and then cause them problem right or take too much time.

 **Karthik Raghu** 14:26  
So.

 **Speaker 1** 14:27  
So I want to make changes to the production only after evaluating anything, right?  
And I I think that is a good idea.  
Let's focus on the US only because it didn't.  
We didn't, even with the scheduling, we did a lot of cost saving, right?

 **Karthik Raghu** 14:41  
Yeah.  
Yeah.  
So that's what right, like, whatever optimize.  
Uh optimization we can do in current.  
Area is it like current state as it is, I think we would be saving a lot of lot of you know dollars if we start doing it now right?  
Like maybe if you can see what happened to POC, right?

 **Speaker 1** 15:02  
Chuck.

 **Karthik Raghu** 15:05  
Like within a week, I think we save a couple of couple thousand $2000, right?

 **Speaker 1** 15:05  
Hmm.

 **Karthik Raghu** 15:10  
So that's my concern here.  
So obviously I'm just thinking like, you know, if we do the same thing for US, maybe probably, yeah, it will be even more for a week.

 **Speaker 1** 15:20  
So yeah, I I completely agree with you.  
I think the only thing why I'm taking time for this one is because on POC, even if we slow down anything, no clients will, you know, complain on production.  
We have to be careful, right?

 **Karthik Raghu** 15:33  
Yeah.  
Yeah, it's very.

 **Speaker 1** 15:34  
Yeah, that that is the only concern I have.  
That's why I'm taking time.

 **Karthik Raghu** 15:36  
That's OK.  
Definitely.

 **Tuhin** 15:38  
Yeah, yeah.

 **Karthik Raghu** 15:38  
Yeah.  
So we don't to have any issues with existing, uh.

 **Anand sivaraman** 15:41  
So, so so Karthik, Karthik.  
One thought that I have is you are absolutely right that we have to start implementing this in production in the US so to the point that Priyanka mentioned, we can look at all the necessary configurations that won't unnecessarily hinder production, right?  
And those particular configurations, if we make changes accordingly and APC cost improvements, we will be able to readily implement it, Karthik.  
So we'll also do one level of analysis and testing to make, they say, testing at least to double check all of this before we move into production.  
So that was one of the aspects.  
So I I Co Priyanka sentiments.  
We will also make sure that we start implementing this sooner, Karthik SharePoint.

 **Karthik Raghu** 16:23  
Thank you.  
Thank you.  
Thank you.  
Thank you so much.

 **Speaker 1** 16:29  
So any other questions or should I proceed?  
OK, I'll take it as it and please feel free to, you know interrupt and let me know wherever you have questions. OK.  
Uh.  
Is I have.  
Do you have a lot of things to cover?  
So I wanted to go quickly, but we can have a discussion wherever you need.  
So this is the same changes that we made for POC.  
This is an Excel sheet right that we have shared earlier with the offshore team as well.  
One point I have not covered in this, but one thing I would like to raise is can we have a call?  
Is it possible to have a place where we can put all these and if we are making changes if offshore team is making changes, DevOps if they're configuring something we can put it so that we know what changed on what date right?  
That way it'll be helpful for us to keep track of that.  
So I know that offshore team also made those changes and might have put them in the sheet, but we we have no way unless we share it on the mail.

 **Carl Droste** 17:30  
Umm, well, OK, so you're talking about which the spreadsheet that you guys were using previous.

 **Speaker 1** 17:35  
Yes, yes, yes.

 **Carl Droste** 17:36  
Yeah, if we want, I mean we could put that.  
You could we could put it into the the team space as well, or we could share the link to it if it if it's.

 **Speaker 1** 17:42  
Yeah.

 **Carl Droste** 17:46  
If you guys own it and we share it, we could share it and put the link in the space and just say hey, uh or we can or we can share a copy and put it into the into the space itself into the files, right.

 **Speaker 1** 17:58  
Hmm.

 **Carl Droste** 18:02  
Uh, this at the team space.  
We could either do either umm.

 **Speaker 1** 18:08  
If anything is fine with me the well, I just think that there should be one common sheet we should be updating and it should real time.

 **Carl Droste** 18:13  
Yes.

 **Speaker 1** 18:14  
We should all be putting in and seeing so that before going in, OK, this was the change made.

 **Carl Droste** 18:16  
Right.

 **Speaker 1** 18:19  
Let's not revert it back or let's discuss kind of like.

 **Carl Droste** 18:20  
Right.  
Well, I I I completely agree that we've got to, we've gotta maintain documentation documenting everything that we make a change for as we go, as we go.  
And if we if we, if we get out of that habit, then I'm afraid things will start to to fall down after you guys are gone.  
And not not being able to give your input continuously like you're able to do right now and things may solely get out of sync again for us and we we'll have these type of problems that'll occur a year from now.

 **Speaker 1** 18:59  
Absolutely so.

 **Carl Droste** 19:01  
It's kind of a mindset.  
So.  
So you let me let me know if we don't.  
How you, I mean we could just share the sheet and you guys can own it at some point.  
We just you hand it off to us and then we take it forward and try to maintain it ourselves.

 **Speaker 1** 19:16  
Sure.

 **Carl Droste** 19:17  
However, you wanna do that, but I agree.

 **Anand sivaraman** 19:22  
That can we not?

 **Speaker 1** 19:22  
You have this.

 **Anand sivaraman** 19:22  
What can we not use Monday for the sprinkler?  
Because every Monday is one place where we store.

 **Speaker 1** 19:28  
So.  
So on and we are putting it there.

 **Carl Droste** 19:31  
Umm.

 **Speaker 1** 19:32  
Actually we are sharing it there, but for for example this kind of sheet the changes even the offshore team is making and they are not on the same.  
I don't know if the if they are on the same, you know pages they have access and they can make the changes because what happened on Thursday is yesterday and day before yesterday actually.  
That option team also, you know, decided that, OK, after all the things that they can also reduce the size of SQL warehouse and that was I think a good step.  
But we have to document that.  
So we share this Excel sheet with them on and, but we didn't.

 **Carl Droste** 20:03  
Truck.

 **Speaker 1** 20:05  
We do not have what changes they make to come forward to see right?  
What changes were there that might have impacted anything?  
So I'm not sure about the Monday before Twain.  
You will be a better person to tell if that is an option.

 **Tuhin** 20:20  
Uh, let me let me take that as an action item.

 **Speaker 1** 20:23  
Sure.

 **Tuhin** 20:24  
Priyanka and I'll.  
I'll.  
I'll come back to you and sensually, umm, to paraphrase right, because I'm not gonna go into as much detail with that team, but just to just to say it back to you, right we are we are looking for a single source of truth log, change log, whatever you wanna call it from both Kanini team and Annex Cloud team.

 **Speaker 1** 20:39  
Umm.

 **Carl Droste** 20:44  
Yeah.

 **Tuhin** 20:46  
Did I understand that correctly?  
Whether that's code infra, what have you, we need a change log common change log for both team to look at.

 **Speaker 1** 20:48  
Yes.

 **Tuhin** 20:54  
Yep.

 **Speaker 1** 20:55  
Exactly.

 **Tuhin** 20:56  
OK.  
OK, good that I can do.  
Alright, let me let me take that as an action item.

 **Speaker 1** 21:01  
So we can come back to that on Monday.

 **Tuhin** 21:01  
Yeah, yeah, yes, yes.

 **Speaker 1** 21:04  
It's.  
So what we showed you on Wednesday was because we were doing it on the 11th and 11th also it was not complete.  
Now it's a complete comparison.  
You can see on your screen from Feb 29 to March 6th and March 7 to March 13th.

 **Tuhin** 21:15  
Yep.

 **Speaker 1** 21:19  
March 7th is when we made the change half day through right and and considering all the days we have saved about $2000 by these changes and there is a distribution also this it is in the Excel sheet based on the services.  
So when you wanted earlier where it's going in which group and these two resource group, as I mentioned last time, these are the ones that are run by databricks, whether it's the resource group which we directly get access to or the data breaks, you know related what data breaks itself creates a resource group and bring up the VM machine, then use the storage.  
Also, there's a sheet where you know the same day to day costing how much it was costing for these resource groups and total ones.  
Make too much 38 to March 13th.  
It's the same thing comparison $2000, which was $4000 earlier, so I'll click if you if anybody has one, let me know as we're over these because we went through these on Wednesday.  
So this is the entire this is what I was putting in the what I put in the summary right.  
We see savings starting from 163 kept increasing and now we almost like Wednesday.  
We had a saving of $500.00 a day.  
Total went up to 2000.  
So to help on that side POC, we are clear because we went through it earlier.  
Any any questions and again.

 **Anand sivaraman** 22:47  
So to break if if you can send shed some light on the fact that these are purely based on configurations, right, we have not yet touched the code configurations and scheduling is what we have done.

 **Tuhin** 22:48  
No.

 **Speaker 1** 22:56  
Yes.

 **Anand sivaraman** 22:59  
So in in I think we have documented the the root cause analysis and what changes and what configurations have been made as well.  
That will be that will be saved and shared with the next load team so that like you know, we can I I think we should also look at that as a best practice if we can suggest something so that any new jobs that get added should also probably follow a similar configuration, right.  
So I I think that would that is part of the scope and that's where we should be going as well, yeah.

 **Speaker 1** 23:29  
Yes.  
And and and that that is the document that we were talking about the the, the table that you see that is coming from that document and we shared it with offshore team to do the same like it has documented everything what was before and what is after and what is the standard practice right, we should be doing even on the job side, right, the scheduling I mean so I'll go for.

 **Anand sivaraman** 23:40  
Yeah.  
Sure, sure. Sure.  
Yeah.  
And and and during during call and Karthik one maybe I'm maybe I'm.  
I'm just jumping the gun here.  
Please forgive me for that.  
What I was also thinking is, you know, once we see certain good uh ohh.  
You know decrease in cost and improvement, improve improvement in the performance and we are able to implement some of these configurations like Karthik was suggesting next week.  
Would it be possible for us to even to to put a proper best practice document?  
Would it be possible for us to probably share these findings with the data bricks team so that we have a discussion then then create this as a you know, tenants of what all needs to be done when it comes to configuration, right?

 **Speaker 1** 24:32  
Hmm.

 **Anand sivaraman** 24:34  
So I was thinking about that I'm.  
I'm just throwing it to you a little ahead of the curve, but that's something that we can definitely do and you know, make this happen as well during and call Karthik.

 **Tuhin** 24:44  
Yeah, I think I think absolutely right.  
Ongoing, ongoing that makes sense.

 **Carl Droste** 24:47  
Yeah.

 **Tuhin** 24:50  
And again, like admittedly right, I think you have heard me say this.  
You have heard Matt say this right?  
Obviously we don't have a lot of domain right in this area, so any best practices you suggest for an ongoing and then think Carly, you're mentioning the same thing like that would be much appreciated.

 **Carl Droste** 25:04  
Yes, yes, yes. We need to.  
We need to communicate everything that we're doing and keep track of it.  
And and I don't know how you would make sure it's one of those things where you almost need to.  
Well, one thing to establish the process since you get into the habit of doing things a certain way so that you track what we're talking about with a change log.  
But the other thing is just, uh, what you're saying is, like, more standardized, you know, understanding exactly what kind of things are good or best practices so that we can capture that and make sure that there's a record of it so that anybody going forward knows exactly how to engage data bricks and manage it so that anybody can can make the performance continue and not step into the same mistakes that we've done before.

 **Speaker 1** 25:43  
Umm.

 **Carl Droste** 26:06  
Agree.

 **Tuhin** 26:07  
Yep, Yep.

 **Speaker 1** 26:08  
Exactly.

 **Carl Droste** 26:09  
Yep.

 **Speaker 1** 26:11  
So the on production we just made, but I'll just say we just made changes and we made sure that you know definitely with the help of the Annex Cloud offshore team, they cleaned up the jobs that were not useful, we stopped them and only these three were running because they were the major ones like they are the major workflow and they had sequenced 3 jobs, right.  
And because because of the timing and I think you might have seen the discussion yesterday like we made them every hour and every two hour because US one was running for one hour, 44 minutes, the other ones like these were running for 45 minutes and it was increasing in U specially because of the ohh you know how did Sir running and I think they might be have been some changes there since Tuesday.  
Uh, they have kept this place open because we were waiting for SQL warehouse changes that the team did.  
I can show you that that they have made certain two of them, at least one on POC and one on EU.  
A smaller for Hertz, for example, the size of the cluster that they're using, and I think they're all aware that, uh, the costing might be because Harrods has some requirement that as sort of mentioned and they have started, they created a duplicate job so that it can just run for hybrids.  
Uh, just wanted to.

 **Tuhin** 27:30  
Yeah, and no.  
I I'm I'm out of it.  
Thank you for calling this out.  
And and and.  
This will kind of make sense to you.  
This is part of the other discussion that I kind of through over to you in regards to connectors, right?  
This is essentially what's leading to this.  
Like you know, it's not directly related, but it's coming in from that point of view.  
The requirement was essentially, you know we needed because we don't have CDCM we need the customers to be able to get access to certain data a from our system, so that that they could build out an integration.

 **Speaker 1** 27:55  
Umm.

 **Tuhin** 28:13  
The solution we gave them was an API and that they could call it every however many minutes from databricks to pull that data.  
But that's where that connector discussion came from and and we don't need to go into deep right now, but just wanted to mention that like that's that's what you're saying here, yeah.

 **Speaker 1** 28:32  
I think, uh, what do you think?  
And if we have time after this, everybody's available, we can talk about these things.

 **Tuhin** 28:36  
Yeah.

 **Speaker 1** 28:39  
Or you want to take it on?

 **Anand sivaraman** 28:40  
Sure, absolutely.  
No, absolutely prank tune.  
Got your point, Sir?

 **Tuhin** 28:44  
Yep.

 **Anand sivaraman** 28:44  
And and I think that calls for a a separate discussion because I I, yeah, we could, we could definitely do that.

 **Tuhin** 28:47  
The separate, separate discussion?  
Yep, Yep.

 **Anand sivaraman** 28:51  
And Prashant and I have compared multiple different integration tools.  
Again, we don't need to necessarily go into the tool aspect of it, but the strategy towards it we can, we can always talk separately with I can set schedule a call at the time that you are all available at you.

 **Tuhin** 28:58  
Yeah.  
Sure.  
Yeah, yeah.  
Yep.  
Yep.  
Sounds good.  
Yeah.  
Yeah.  
Yeah, that's that's why the jobs are running as fast, right?  
It's a little bit of a our our doing.

 **Anand sivaraman** 29:13  
Sure.

 **Tuhin** 29:15  
So we'll have to unwind that.  
But you know, anyhow, I understood pain for for this exercise.  
Right.  
Who?  
I you cannot be optimized.

 **Speaker 1** 29:22  
Yes. Yeah.

 **Tuhin** 29:22  
Optimized, right? Yeah.

 **Speaker 1** 29:25  
So the first we have the AP SE, you can see the difference of almost $1000 for rescheduling.  
I think that is what Karthik was just the job.  
You know, setting up the jobs in a way that they don't fail continuously.  
That's all we did and you can see the difference of around $1000 in a week, right?  
And there's also another thing in that we notice is in all environments you'll see forward also and you might have seen it that as I mentioned in the two weeks ago or such when I started looking at that, that you have a bell curve if day wise it was changing, but you will start noticing there's a constant you know after doing those changes, scheduling properly, there's a constant a cost right, like for example 100 under $100, this is API.  
I know there's less data and everything, but it is keeping under $100 only where it was $200.00 and it was varying.  
There's no there was no correlation apart from weekdays.  
That's just the one observation that we did and the same the same costings for everything you can see there was $1830 cost and especially the only changes I see is that what we may changes was which should affect this the cost for virtual machines while job scheduling and it went from $600 to $200.00 in the next week when we reschedule them and even the Azure Databricks crossed went down from you know $800 to to 100.  
So that was like 1/4 out of one, a little more than that.  
So.  
So the total cost that we saved on AP for by rescheduling, you can see the savings and all they comparisons above 100 to 200.  
And I see a pattern that that it keeps increasing as the day goes right?  
Ohh so this is for AP Southwest US E There is changes but yeah there we can see quite number of you know variation.  
I won't say the same for her, but yes, we did save $1000 again here by just the rescheduling.  
And so it went from 4398, which is there, and you see there's your database cost and virtual machine cost.  
As I said, that is the ones that should be.  
It went to hound half for both from 1900 to 1300 and a virtual machine.  
It went down to 500, so overall cost you can see this $100.  
This is one thing that I missed on the first summary.  
I'll connect that before sharing in EU.  
Umm, it did go down.  
Uh, but then it went up again for the last couple of days, so the past is if almost the same, it's not much difference.  
It's $60.00 about the the.  
Not as I think that's all for the cost.  
One thing I wanted to share, as I said this is the screenshot that I when I started noticing this forecast change right.  
As I said, it was $82,000 on the start of the March when we looked at it when we started working before March 7th.  
Then on March 13th, we saw the forecast changed to 75,000 and today this has gone down to 67,000.  
I'm hoping it'll go down more as the days go by, so any questions on the cost?

 **Tuhin** 32:47  
No, I think I think that's good for now and we'll like take away is that I think we'll we'll as Karthik was asking like you know, we'll have to think about the production stuff.

 **Speaker 1** 32:47  
In this.  
Hmm.

 **Tuhin** 32:57  
We'll we'll think we'll we'll figure out a strategy, maybe even if we cannot optimize EU, we can optimize.

 **Speaker 1** 33:00  
Umm.

 **Tuhin** 33:03  
I think the one that decision will come down to right, maybe AP is low risk for us, but we'll confirm that maybe we go ahead and do that there.

 **Speaker 1** 33:09  
Hmm.

 **Tuhin** 33:12  
Then we'll do the US and and move on and we can we can talk about that data.

 **Speaker 1** 33:16  
Umm.

 **Tuhin** 33:17  
That's good vacation, right?  
Because you're strapped for time a little bit right now, so let's just go over that.

 **Speaker 1** 33:20  
Sure.  
Yeah, I yeah.  
So I think you might have all seen this slide as well.  
There is 2/3 approaches actually that we have covered.  
One is the one that when you were mentioning refreshing the dev with the prod data which no optimization required.  
So we'll only activities that we have to do is we'll work with the next team DevOps team to validate any existing scripts that can load the data from prod to Dev or if they don't exist.  
So we can create them right?  
And then the DevOps seem can run the scripts in production and download the data.  
Can any will verify in Dev that it is there and prod data is validated?  
We will make sure with the team, right.  
The assumptions here are that continuity team is to be given access to the prod data only in dev environment and also that annex team will manage the scripts and run in the production.  
No need.  
Pros are no need for scrubbing.  
Data becomes very easy available for undresses and you know so also less time right there.

 **Tuhin** 34:23  
Umm.

 **Speaker 1** 34:24  
It is going to take you one business day.  
Cons are access to production data that can in in team will have and it'll be non prod.  
Another calls.  
You are already aware about legal bindings.  
If you have for your company right for companies data.

 **Tuhin** 34:39  
Yeah, and yeah, OK, got it. Understood.

 **Speaker 1** 34:39  
Umm.

 **Tuhin** 34:42  
Thank you. Yeah.

 **Speaker 1** 34:44  
So approach to is to anonymize the prod data and hair again the community will validate the if existing strip, otherwise we'll change it.  
Ohh, I think it got uh duplicated.  
I'm so sorry for that.  
So the second one that was there, we'll validate the scripts and we'll download the data.  
Then we will actually anonymize the data in non prod.  
So again, that'll be scrubbing.  
Will be required.  
Please ignore the you know no, uh.  
Details on the screen because that got copied from the the approach first, but yeah, just to give you an overview, it'll be pros will be that you will have anonymized prod data in the lower environment.  
You will have, you know you can use it for cost testing and performance testing as well as any issues that can arise with the data that will be there, right.  
And you will have to data cataloging and this is for the long term.  
It will be good process also to you know to make sure we are following the right process and nobody apart from the business is accessing the especially the vendors are not accessing the prod data.  
That is also what would be good cons are that it'll take time for you to bring that down.  
It'll be one time exercise that will take some weeks.  
So and the timeline that we have put is 2 weeks.  
That should be overall and the assumption here would be that coming the Community team can help, but it'll be a Nextcloud team that will be providing and classifying the data and identifying what needs to be anonymized.  
Anonymization scripts will lie with responsibility with live with colony and we can provide that to DevOps team to run that and so that we can directly have access to the scrap data in the lower environment.  
Uh, approach three is another one that we discussed, I think on Wednesday.  
The in that is the like generating the synthetic data.

 **Tuhin** 36:39  
Yep.

 **Speaker 1** 36:39  
Umm, so you're beginning to solve the, you know, we can solve the problem of.  
Pre, uh having the data?  
Uh, broad data in, you know, and we won't have to scrub the data.  
We can generate the data or using the scripts and data will become again the same thing as scrubbing the data.  
It will be the similar kind of data.  
It'll solve the performance and cost issues.  
We can long term, we can do that and also one more thing is that nobody's accessing your prod data.  
The activities will be that we will, you know, still there will be some time because we will have to make the scripts according to what kind of data should be.  
Uh produce and there should be a seeding of the data here, right?  
Which is similar to production data that should be there and then we can, but this all can be done on non prod environment. Ohh.

 **Tuhin** 37:29  
So Priyanka one quick question understood on the approach three is, is there like a kind of a tool you are using or you you'll write the script to generate the synthetic data or is there like a because is there a cost implication with like buying because I know there are some tools out there that lets you do that.

 **Speaker 1** 37:33  
Umm.  
He he.

 **Tuhin** 37:46  
So what's the approach?

 **Anand sivaraman** 37:46  
So.

 **Tuhin** 37:47  
Yeah.

 **Anand sivaraman** 37:48  
So do you and I'll, I'll take the I'll take a stab at it, because this is something that we have done just one week back with another client.

 **Tuhin** 37:48  
Yep, Yep.  
Yep.

 **Anand sivaraman** 37:54  
So all the tools that we are going to use are open source tools tune.

 **Tuhin** 37:54  
Yep.

 **Anand sivaraman** 37:57  
There are Thora of tools in the Python ecosystem that not only allows you to anonymize the data, but also helps in synthetic data.

 **Tuhin** 37:57  
OK.  
OK.

 **Anand sivaraman** 38:07  
So we we have explored a few options but we want it to be sure because see, it's not just about generating the synthetic data right.  
It's about making sure that the data makes at least some sense.

 **Tuhin** 38:17  
Yeah, right.

 **Anand sivaraman** 38:18  
If not every too a whole lot, so that is the reason why we took the Python route because it at least allows us to do some customization.

 **Tuhin** 38:19  
Right, right. Yeah.

 **Speaker 1** 38:19  
Umm.

 **Anand sivaraman** 38:26  
If we had to go with the established market established tool like Dell Fix or whatever it is going to cost and it's also going to take time for us to implement you.  
And so that's the reason why we are taking this approach.

 **Tuhin** 38:35  
Yep.

 **Anand sivaraman** 38:37  
That's the only reason, yeah.

 **Tuhin** 38:38  
Makes sense.  
Makes sense.  
Thank you.  
I just want to double check, right, because there's also like managed services out there, right?  
Also wasn't sure what what are planning on doing, yeah.

 **Speaker 1** 38:44  
Yes.  
Yeah, yeah, actually for the this because you have so short that line right now and short short timeline, we have actually, as Anand said, we were considering only the you know open source or something that we can provide and we can build so that we can have control over the timeline as you said the other products they are there for even the scrubbing of data.

 **Tuhin** 38:51  
Umm.  
Mm-hmm.

 **Speaker 1** 39:06  
But you have to get engaged with the vendors and that will take time itself.

 **Tuhin** 39:08  
Umm yeah.  
Yeah, exactly.

 **Karthik Raghu** 39:12  
So no one question I have like you know, not scrubbing the data.

 **Tuhin** 39:12  
Makes sense.

 **Karthik Raghu** 39:16  
So why do we need access to education environment?

 **Speaker 1** 39:22  
So this is for the seating of the data.  
The data should be similar to production data, right?  
So at certain point we will have to do that, but afterwards it won't be the same, right?  
It's a limited time access and after the process is set up then we don't.

 **Karthik Raghu** 39:36  
Anything, just to make sure that you are, uh, like you need some template to.

 **Anand sivaraman** 39:43  
Yeah, yeah, yes, Karthik.

 **Speaker 1** 39:43  
Yes.

 **Karthik Raghu** 39:43  
Create data so.

 **Anand sivaraman** 39:44  
So so there are two aspects to the template Karthik.  
One is the schema of the production tables, Kartik because many a times, we don't know if the production schema and the lower environment schema are same.  
So that is 1 aspect.  
2nd is some sample data, even if it is stale data it is fine.

 **Karthik Raghu** 39:56  
OK.

 **Anand sivaraman** 39:59  
But the data needs we need to have some sample data so that what when we do this synthesization right there are it's not always a rule based approach.  
Karthik, many of these Python libraries.  
They also have embedded machine learning algorithms, so typically we use algorithms like, you know, K nearest neighbors and and algorithms that are able to find similar or close by data and then synthesize the data based on that.  
So there are several algorithms that are available, so for that we need some seed file, some base file.  
It may not be actual production file, even if there's something sample one year old, 2 year old data that's also fine.  
So that like we can use that to start synthesizing the data thing.  
That's the only reason why we need that you are absolutely correct.  
We typically don't need production access, but because we are doing it for the first time, we have to do it.  
And if you see that is the reason why we have put an assumption that you know while we will write the script, the running of the script might have to be done from from unexcused perspective because we won't have access to run this in production Karthik.  
So that is the only reason we wanted to be sure that you know it doesn't touch all of the you know it it is.  
It is making sure that the production data is sacrosanct and you know we are clear about what our protocols are.

 **Karthik Raghu** 41:15  
So in case if the if anything changes on the schema side right?  
So like, should we actually look at changing the script to like in future?

 **Anand sivaraman** 41:24  
In in a?  
Yes, Sir.  
Yes, Sir.  
We have two so so that is a very good question.  
Thanks for asking that.  
Ideally, if we want to do this as a repetitive exercise for future right, we will have to create an anonymization per pipeline or a synthetic data pipeline which can be run on demand Karthik.  
And as part of the pipeline, we can actually embed schema evolution to tell you the truth, we can use the same data bricks environment to do this synthesization as well Karthik, if you ask me, I can write a Jupiter notebook, run the Python code within Jupiter notebook, use data bricks, CDC and schema Evolution option which is part of the databricks environment to do all of this.  
We didn't want to use databricks because it is adding to the cost.  
That's the only reason why we are running this as a separate exercise on a hosted.  
You know, just on a Python or to put a notebook hosted on a particular server.  
That's all we have, but you are you are absolutely correct.  
If you want to strategize this in a regular basis, we will need a proper pipeline which will also automatically understand if there are schema changes and appropriately do this Karthik, we can always work on that Sir.

 **Karthik Raghu** 42:28  
OK.  
Yeah.  
Thank you.

 **Tuhin** 42:33  
OK, sounds good.  
Given that we are overtime pianka did is, was there anything you wanted to OK, there is duplications.  
OK, let's run through that then.

 **Speaker 1** 42:41  
Yes.

 **Tuhin** 42:44  
Then we'll we'll recap the eyes.  
Yeah, yeah, correct.

 **Speaker 1** 42:47  
Sure.  
So as I mentioned, we just put couple of them here.  
One was that you know your updates whenever they are happening.  
It's in the code.  
The way the code is dealing with the data right now, when it is ingesting is you have one record and it has a certain update date.  
Because we're running between a timeline, it just looks at the update date and get another, you know record for the same everything same just you know maybe your transaction data is different but it does not.  
It makes it as a new record in the bronze layer, which was, you know, just an updated record in the previous one.  
So you can see the example like if the credit was zero and now the resistant, but it'll now take as if you have two records for the same transaction ID, right?  
And that that is one reason you're seeing duplicates in transactions.  
Same the deleted records in SQL are not counted while ingesting as they are not deleted from the bronze layer.  
So you might not find that record in DB, but you will see that record still being there throughout in the reports because they are in the bronze like currently nothing.  
We can skip the stores one for now and that is we can go in detail later so.

 **Tuhin** 43:50  
OK.  
OK.  
Sounds good.  
Then the AI's what I understood and you know, please correct me if it's wrong. Right?

 **Speaker 1** 43:59  
Hmm.  
Hmm.

 **Tuhin** 44:01  
So and from Annex Cloud side, as our what we need to implement a chain change log?

 **Speaker 1** 44:09  
Umm.

 **Tuhin** 44:09  
Between Kanye and and Annex Cloud database team, the second thing is that you need a decision on the three approaches.

 **Speaker 1** 44:14  
Umm.

 **Tuhin** 44:17  
That has been, I suggested by the Community team and pick one for the changes.

 **Speaker 1** 44:20  
Umm.

 **Tuhin** 44:23  
Uh, and the third AI from from our end that we need to let you know is that which environments can we start changing the configuration that we have done in POC.

 **Speaker 1** 44:35  
Umm.

 **Tuhin** 44:37  
Tentatively, we are saying that it is probably going to be a P first, then the US later, but I'll confirm you.

 **Speaker 1** 44:42  
Umm.

 **Tuhin** 44:44  
So those are the three things I took away.  
Uh, yeah.  
Is there anything I missed from our end? Yep.

 **Speaker 1** 44:50  
One thing, uh, yeah, for the duplicates, as I mentioned earlier, right, for duplicates missing data.  
We have certain issues that we have identified.

 **Tuhin** 44:56  
Yep.

 **Speaker 1** 44:58  
So we did discuss how do you want to take the new approach?  
Do you wanna write a code starting fresh because it's a different kind of dealing, right?  
It's right now the code is dealing with incremental data and we want either the delta data or we want the full load right.

 **Tuhin** 45:08  
Umm.

 **Speaker 1** 45:13  
But that is both new code, right?

 **Tuhin** 45:14  
OK.

 **Speaker 1** 45:16  
So you said you will come back to us about that approach also how we want to start slipping if that is the case?

 **Tuhin** 45:17  
OK.  
Yep, absolutely.  
Yeah.  
And Rupali, if you if you could just please also like clearly call out the AI as a separately between a Canadian annex cloud.  
So we we keep track of it in a minute.  
That would be much appreciated. Yep.

 **Rupali Ghosh** 45:40  
Sure doing.

 **Tuhin** 45:41  
Yep.  
Thank you.  
What I don't have anything else.  
Unless like power Kartik Anand, you guess when I add anything?

 **Carl Droste** 45:51  
No, no, I'm good too.

 **Karthik Raghu** 45:51  
I'm good today.

 **Tuhin** 45:55  
Thank you, Tim.

 **Anand sivaraman** 45:55  
Two in one.

 **Tuhin** 45:56  
Yeah, go ahead.  
Yeah, go ahead.  
Yeah, yeah.

 **Anand sivaraman** 45:57  
Sorry, can SO11 point is I think once we get you your approval on either either ways right either synthetic or scrubbing or just dumping we will we will move forward.  
Meanwhile, I've asked Priyanka and team to start working towards all making all of this, you know, ready so that we can quickly jump into action once we get the approval on it from you to one.

 **Tuhin** 46:24  
Yep.

 **Anand sivaraman** 46:24  
The the email that I sent yesterday summarizes all of this.  
If you have any question again I can get into a call with you and hash out any concerns you may have, Sir.

 **Tuhin** 46:30  
Yeah.  
Yep, sounds good.  
Ah, you know, if I was a betting man, I think Matt's gonna tell us to go with option one.  
But you know, I don't speak for him, so let me just talk to him first then, then come back to you.

 **Anand sivaraman** 46:45  
God, it's.

 **Tuhin** 46:46  
Yeah.  
Alright, sounds good.  
Yeah, alright. OK.  
Thank you, Tim.

 **Speaker 1** 46:53  
Thank you.

 **Carl Droste** 46:54  
Thank you.  
Thank you all.

 **Speaker 1** 46:54  
Have a nice weekend, everyone.  
Bye bye.

 **Carl Droste** 46:56  
Yes, you as well.

 **Tuhin** 46:56  
Thank you.  
Bye bye.

 **Natarajan Sundararajan** 46:57  
Thank you.

 **Anand sivaraman** 46:58  
Thanks, Carl.

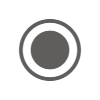
 **Rupali Ghosh** 46:59  
Thank you.

 **Anand sivaraman** 46:59  
Thanks kartik.  
Thanks to you.  
Thanks everyone.

 **Tuhin** 47:01  
Like thank you.

 **Karthik Raghu** 47:01  
Thank something.

 **Carl Droste** 47:02  
Thank you all.  
Very good information.

 **Tuhin** stopped transcription