**1 Sep, 6.32 pm**

**Transcribed by** [**TurboScribe.ai**](https://turboscribe.ai/?ref=docx_export_upsell)**.** [**Go Unlimited**](https://turboscribe.ai/subscribed?ref=docx_export_upsell) **to remove this message.**

No worries. Alright, no worries. Yeah, sure.

Let me know once it is visible. Yeah, I can see. Okay.

So yeah, just let me give you a recap, a brief about what we had discussed so far in the previous sessions. So we had all these data sources on the left side and we have these sizes. So initially I could able to work around these four datasets and this was the challenge for us, these four datasets.

Yeah. So now I could able to manage to handle this data and we have the results for all these green highlighted datasets which includes the browsing, file access, instant messaging and then streaming and wipe. So let me just take you on to the notebook.

Yeah, so in the last results I will have these fields. Some of them I have kept it intentionally so that I can validate the join whether it has joined correctly or not. So here you will see the report timestamp and this is the CGI coming from the reports.

And then on the right side you will see the CGI wipe which is coming from the another dataset from the begin time. So I have just kept it. Sorry, it is coming from the which datasets? Yeah, this is the SOC datasets where we have the begin times and everything, CGI length and all.

Yeah, I mean the... Yeah, yeah, so CGI wipe is from the SOC datasets, CGI X is from Nintendo. Correct, correct. Yeah.

So and then we have the report timestamp and the begin time. Yeah. And we have the identities and one thing that I have noticed here is we have this masked MSSD throughout.

So I am keeping the hash also. Then once you go down, so I just check if we consider all those six datasets then we have around 7.1 million. But this is within the delta of 10 seconds.

And ideally what we should consider is where we have the exact timestamps. So which is around 4.2 million identities. And 1 million, that's a lot.

And so remind me the... Could you show me a sample? Could you show me a sample for how the F3 looks like? This is the sample, yeah. Scroll to the side. Yeah, right side.

We will have this time difference also. Meaning that this data session ID was created within the delta of 10 seconds. Yeah, yeah.

Fully aligned on that one, yeah. Completely understand. Interesting.

So just to be clear, you did the match on the time and on the CGI. Let me show you. So how are you ensuring that you're not picking another customer who for one reason or another was on that CGI at that exact moment? Wasn't the idea for us, I mean, wasn't there another parameter we were looking at? So I think if you aggregate the data you've just identified, and then look at the time diff.

So for the real customer, that time diff should be almost always very small, like almost always zero. But then for the person who is not, he'll have a few records, and it will be scattered. It will be a fairly random distribution.

Okay, so if I go to this time frame, this sample. So let me just try to understand the question that you're asking. So can you just come one more time and explain by just looking at this data frame so that I can understand? So there was one extra dimension that we needed to include, I mean that we were to include to be able to say this is the customer.

So from here, we've done all the big steps. So we've looked at the CGI, and we've looked at the time they were on the CGI. And we've kind of shown that using time diff.

How many seconds, this is seconds, right? I mean minutes. Seconds. How many seconds were they on the cell? Yeah, how many seconds were they, I mean, how many seconds in the future or in the past were they on the cell when this data record was generated on the SOC side? Yeah, so we've done that much.

But now there's one additional thing. What if there's another customer who's on that cell at that exact same time? So at least over here, you can see there's this same customer, 25471427. He has four records.

One, two, three, four, the last four. Yeah, he has those last four records. And you can sort of see.

So this is the same data record. The same data record, but then it seems he has multiple turned out records on the other side. Yeah, probably multiple data turned out records.

Yeah, can see those timestamps there. So I think, so the scenario would be, how many times do you see this customer on the same cell over the seven days? So like over the seven, over the 14 days that you've considered, you found these 4 million customers who exactly match. So how often do they exactly match? There are some customers who only exactly match once.

There are some customers who exactly match all the time. So if you find out the total number of times we see, I mean, the total records we have for them that have matched, even with a delay. And then we find what proportion of this customer's data do we find an exact match? That can be an indicator.

I mean, that can, that will confirm for us. Okay, this is, this person is, this person we don't know is this person. Okay, okay.

So let me just recalibrate what I understood. Let's say we have this many customers only for the exact timestamps, 4.2 million around. But we want to understand out of this 4.2 million customer, what is the distribution of each customers is connecting to the sale at a point in time? Maybe a count of, yeah.

So is that the, yeah. Yeah, it is, it is. So distribution is correct.

So, so it would be distribution of, of time diffs across the 14 days. How does that look like? So if you look at, so for that one customer, if you filter by only them, and then look. So I've noticed that, I mean, I've noticed that you like these multiple, I mean, multiple records are being matched to the same stock record.

So, so for the analysis to be a bit robust. Maybe you might need to just keep the one Tandau record that matches closest to, to the SOC record. Like, like a closest match, like the smallest time diff for each for each data record.

So you need to do that fast because I've not had these multiple Tandau records for, for the same SOC record. Yeah. That's what I think.

Yeah, because, because, because for you to create that distribution and for that distribution to represent, like, I mean, to, to represent the, the, the matches, like how close the, the, the proximity or the closeness of a customer to match to a specific data record. So the likelihood that a Tandau customer is matching the SOC data record, for you to be confident that that's the person, you need to, I mean, you at least need one record per SOC data record, I think. What do you think? Have you understood what I'm saying? Yeah, I have understood.

And I want to add on that is if, if we just look at this, we are already vanishing those duplicates, right? So, won't that be enough? Because if we do the study that you recommended, then how and what answer we are going to give? Again, it's a unique identities only, right? Yeah, yeah. Do you have that initial image? Remember when, when you and Rania first joined the team, there's an image I created for you and sent to you. Yeah, I have it.

You still have it? Yeah. Could you share the screen? Yeah. Yeah, that.

So here we are. Yeah. I can see it.

Yes, yes. So we are using SOC as a fingerprint for a given customer. That's, that's the idea.

So over here, so SOC everyone, we have like everyone who's connecting to some data service on SOC. Then every cell, everyone connects to. Then in Tandau, we have the cells that its users connect to at a given time.

So, finding a customer isn't, when I talk about fingerprints, it's, it's not like one record to one record, like neither is it one exact match. So yeah, it's good you've gotten an exact match. But, so what gives, what will give us more confidence is that they've been multiple exact matches over, over time.

So if you look at 14 days or even five days, and by the way, you don't even like, as I mentioned initially, you don't need other for this. All we're looking for are close matches or exact matches. If you look for the sequence of cells a person has connected to, and they match with what, what's on Tandau, very, very, very closely, then we can be confident that, okay, this is the person.

So that's what fingerprints here means. Yeah, so like, you can use one MSSD, one CGI, that's an exact match. But the problem with that is that what the problem I'm raising, someone else, because there are millions of customers over a single moment, so there are hundreds of thousands of customers who are on a given cell at a moment in time.

Actually, no, sorry. There is maybe 30 to 50 people at a given moment who are on one cell. And each of that person could be using data at that moment in time.

So an exact match might not give you, I mean, the fact that it's an exact match is, is very, very compelling. Because, like the likelihood of other customers sending data at that exact moment is, is low. But the problem is, there are also maybe 25 to 30 other people who are on that exact cell at that exact moment pushing data at that exact time.

And so the only way you can eliminate us identifying those people as, as, as the user is by looking at, okay, looking at now, two, three days prior, and then finding the people who keep getting consistent, consistent, exact matches. So like two, three, five exact matches. And that's what I mean by fingerprints.

Yeah, it's making sense. And to validate your thought, I'm just thinking if I just go back to this, this sample. So for the report data, I have a report timestamp, but not necessarily, if someone can, can you hear me? I can hear you, but your screen has gone black.

Yeah, I'm not sure what happened because it got blank. Okay, so let it be. So what I was talking about the data report timestamp that I had.

So let's say if I have connected to some of the nearest cell, and there may be n number of people connecting to the same cell at a given point in time. Then I have a unique report data session. Okay, okay.

Okay, I understood. So basically, if n number of people are connecting to that cell at the same time, then if I just find out the number of times that particular person is connected to that cell, that can give us a confident in saying that, okay, this is the exact person who connects to that cell. Yeah, exactly.

The only difference is, it's not just, it's not that specifically, it's, it can be any number of cells, but that they have exact matches with, with multiple other cells, when they are pushing data on and turned out. Yeah, or very, or very small delay times. Yeah.

Okay, so, yeah, I will do it on. First of all, one example only. And then maybe we can Yeah, yeah.

Yeah, actually, actually, what what what I suggest is confirm what we are saying. So like I'm effectively I'm saying, it's very likely you'll find someone else who has an exact match. So, so just check if you have more than one exact match on on the same cell at the exact same time.

Yeah, but but even so to just check that and then once you've confirmed that that's actually the case, yeah, you should now just focus on just one person and look at multiple cells. Yeah. Yeah, I will do that.

Okay. Yeah, great progress. Glad to see at least our first part is done.

That's the, that's the most, that's the most challenging part. Yeah. And so the second part is much easier.

So like the expected my expectation from this is for you to create a data set. So that's what I'd like you to do. So we sample data you've worked on just the 14 days, that's enough.

Just create a data set, and I'll tell you where to dump it, where to put it on S3, because it will be very useful. Even as it is, as in that form, it will be useful for for network insights, it will be useful for Jeff, I mean, it will be useful to the team. So yeah.

Okay, so it's already into S3. But yeah, I will perform the next session and I will show you the location as well. Alright, cool.

Awesome. Anything, anything else? Yeah, I wanted to show you the other task progress also, but I'm not sure my screen is freeze completely. I don't see any option.

Yeah, so maybe unshare and share. Just a second. Because on my screen, I am completely blank.

Let me just drop down everything. No, I'm unable to even open the video also. Yeah, but just give them and just let me give you one highlight on the another task.

So I just spoke to Mary in the morning, if I can connect with her regarding the inference that we are trying to get from the aggregated IVR table. So there were two questions like how we can infer those subscribers who have received the offer and they called to the service center. And those who were already aware and did not call to the service center.

That's the inference, which is going to set the meaning for the conversion rate for us to measure on the June based data that we have received from Mary answering. So I have prepared some of the document to show her so that she will get familiar with it. So I will just walk her through initially on my understanding on the IVR process completely.

And then we'll see if she can able to help us in inferring those subscriber. So let me repeat, let me just repeat what you said, just so I know we're on the same page. So you need the people who didn't know about the self-service option, like I've never experienced it, but called the call center.

And those who did know about the self-service option and didn't call the call center. Is that correct? Yes, correct. Okay.

So, so I'd imagine, is it possible to, is it possible to, to like just look at the historical data, like look back two years or three years, two years within the data and use that as the knowledge of self-service. Yes, that is possible. But when you look into the, so basically there are two tables.

The first one is the campaign history table, which tells you the history about all the campaigns, historical campaigns. The second table, it tells you about the option that particular subscriber has used. Okay.

And those options are related to your campaigns. I mean, that could be your service offer or anything other. So I have that also in my Excel sheet.

I could have shown you, but let me give you one idea. So there is a, there is a majority of check swap pin option where the majority of subscriber is using that from the IVR table. So when any particular campaign is being run and the items which is tagged to that campaign is as a seem swap or maybe some fraud related item.

Then how we are going to know that this particular subscriber has used this particular item under that option selected. So I could have shown you the entire process. Or maybe shall I drop for a minute and then I can rejoin.

Yeah, sure. Yeah, just give me a minute. Okay.

Can you hear me? I can hear you. Okay, cool. So let me reshare my screen.

Yeah. Is it visible? Yes. Okay.

So here you see, this is the aggregated IVR self-service table. And here below you will see, this is the another table. So basically the below table records all the process flow of the IVR for a particular subscriber.

For example, when the call started, yeah, every option it is recording. Now this above table is the derived table from this SFC IVR info. So ideally we should refer to this table.

That confirmation we have received already. Now let's say this MSISD. I have just taken one example.

This was involved into the campaign which was run on to specific date. For example, let's say this is the Digicorp student info, right. Campaign code was this and that particular MSISD falls into this targeted group.

Now let's say, so let's say if I want to get the information for that particular MSISD. Okay. So here is that information.

And when I look at these options selected by him or her, then I would be more interested into this option. Check swap without pin. Right.

So now my campaign is saying that this was the student info, something like that. Let me open the email also. So what I am trying to infer here is, so this is the thing that Mary has shared with us.

And these are the items, right. SIM swap, PUK and SIM card related issues. So how we would know from those selected option that this particular parents or subscriber has responded to it.

Whether they have used it or not. And whether they used it, then they really called it or not, which is necessary for this action. Yeah.

Yeah. So I'm going to discuss this with her. So I have this set up call with her tomorrow.

But wouldn't one way to do it be like checking? Do this is only self-service for the IVR? It's only the IVR. What if someone has done self-service from some other portal or channel? Would you have visibility of that?

**This file is longer than 30 minutes.**

[**Go Unlimited**](https://turboscribe.ai/subscribed?ref=docx_export_upsell) **at** [**TurboScribe.ai**](https://turboscribe.ai/?ref=docx_export_upsell) **to transcribe files up to 10 hours long.**