

# Digital Identity Platform: Case Study

Subhra Mukherjee

# PART 1

## Data Ingestion & Validation

- The Raw **CSV** file has the embedded double-quote characters represented by a pair of double-quote characters making the direct ingestion into the BigQuery platform tedious
- I wrote out a small [python script](#) on the Kaggle platform (since it does not require any setups on the local environment and the data is private) to clean the double quotes and put the de-quoted column names in place
- This cleaned CSV is now ingested to create a Table in BigQuery on which all the Data Validation and Cleaning is done
- Most of the Data is for 2024 January & February, there's a few timestamps from 1994 & 2034 which are probably erroneous. There's also a few rows with **kyc\_analyst** as null and **event\_type** is unknown. Excluding these we get the validated KYC data
- The SQL file called **data\_validation.sql** implements these (*please note that BigQuery SQL dialect is slightly different from Snowflake*)

# Assumptions during AHT Calculation for Client Operations team

- Based on the Data i see that any case may be handled by multiple analysts who may or maynot submit results while the case may get reassigned. For example, **case\_id = 56ddc710-1142-41f0-bb7e-2ede76684769**



- I assume the entire time frame from the first assignment to the last submission is the handling time so  $(t5 - t1)$  in this case
- There's about ~5000 case\_ids where there's only 1 event for a case. This would mean a **0** handling time for these cases. These are likely to cause an underestimation of the AHTs so I exclude these when calculating the AHTs

# AHT<sub>(in seconds)</sub> for Client Operations team

	AHT (When the case level final interaction is Submits Results)	Case Count(When the case level final interaction is Submits Results)	AHT (When the case level final interaction is Assign Verification)	Case Count (When the case level final interaction is Assign Verification)	AHT (overall)	Case Count (overall)
Jan-2024	711	222049	61813	195	765	222244
Feb-2024	884	88722	48309	104	939	88826

- Cases which remains unsubmitted, drag the AHT up significantly even though these are very few in numbers.
- Such unsubmitted cases require (on average) 2.1 analyst intervention, while those where results are submitted have ~1 analyst intervention, indicating that these cases are likelier to involve more analysts. A recommendation could be that when the handling time of a case reaches a certain threshold (like the 99th percentile of AHT when results are submitted) then they be immediately transferred to a special queue which is manned by more experienced analysts trained to handle anomalies in the KYC data rather than having different analysts taking a look at the case
- The attached SQL file called **AHT\_Operations\_Team.sql** is used to calculate the AHTs

# Assumptions during AHT Calculation for individual analysts

- As seen for `case_id = 56ddc710-1142-41f0-bb7e-2ede76684769`



- To keep the definition of handling time coherent for individual analysts i calculate handling time as **(last submission - first assignment)** per case. No analyst makes more than 1 submission for a case. So for Analyst 1 ( $t2 - t1$ ) and for Analyst 3 ( $t5 - t4$ ) in this case
- Since Analyst 2 does not submit anything, i don't calculate a handling time for him/her. This specifically because there are cases (e.g. `case_id = 'a4bd07e2-d8e6-4da2-9b3d-07a307a2c034'`) where a particular analyst keeps getting re-assigned a case without submitting
- Also excluding cases where analyst just has a submission for a case which would make the handling time appear 0 for the particular case

# AHT<sub>(in seconds)</sub> for individual analysts

	AHT analyst	Cases handled
0d035d377c9bf245ea9e100bc8e7adfd	23937.08	243
8514a883217c0ef20dfa3eb8214719c5	19909.68	41
e300ce13cd7c9f0a349271458d164cef	13474.32	153
b2cb12b35ec4065ff604bec73094ed81	714.11	4521
a64320e8fb555d5237908e87179d9bd6	316.65	3867

- Top 5 AHTs for the Analysts. Attached CSV **AHT\_per\_analyst** has the entire list for 53 analysts
- The attached SQL file called **AHT\_Analyst.sql** is used to calculate the AHTs

# PART 2

## N Onboarding

### Key Metrics/KPIs to be used to test Hypotheses

**Conversion Rate (user level):** Number of users who completed the onboarding flow as a proportion of all users who landed on the account creation page

**Conversion Rate (Session level):** Number of sessions that completed the onboarding flow as a proportion of all sessions that commenced a session

**Number of Sessions per Verification:** Number of times a user goes through the flow before account opening or eventual drop off

**Number of Rejections per Verification:** Number of times a user is rejected before account opening or eventual drop off

**Dwell Time per stage:** Time spent per stage (of the onboarding flow)

**NPS scores:** NPS scores for users who complete the flow

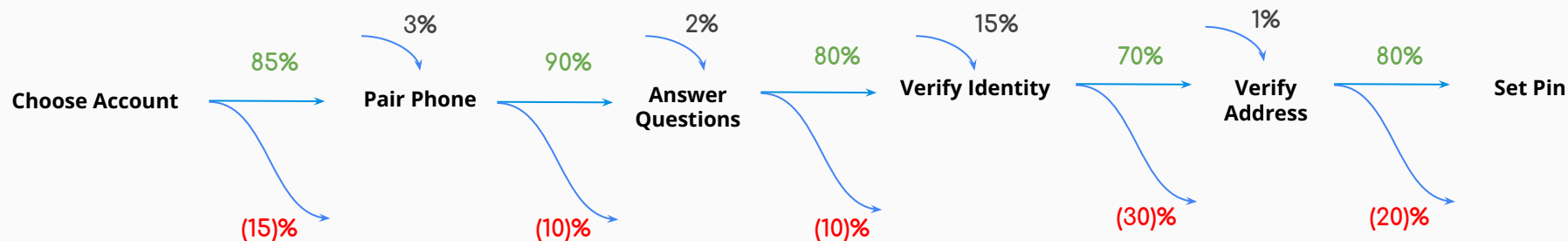
**Fraud detection error:** Ratio of the number of **incorrect** predictions of fraud to the total number of users

- **Fraud detection error (False Positive)** - Ratio of **incorrect** prediction of Fraud (Predicted **Fraud** when actually not Fraud) to the total number of users
- **Fraud detection error (False Negative)** - Ratio of **incorrect** prediction of Fraud (Predicted **Not Fraud** when actually Fraud) to the total number of users

# N Onboarding

## Key Metrics/KPIs to be used to test Hypotheses (continued)

**Flow Visualisation:** A Linear Flow visualisation of the onboarding flow (like the one shown below) simultaneously visualizes multiple key KPIs like:



**Made it Rates (MIRs):** It captures the proportion of Sessions who 'made it' to the next stage of the flow.

**Drop off Rate (DRs):** Complementary to the made it rates it captures the Sessions that dropped off a stage

**Entry Rate:** Sessions that start a session at a particular stage



# Hypothesis 1: Prompts during Selfies/Images/Videos

- Live Prompts when clicking Selfie/Images/Video during identity verification:

Improper selfies/images/videos during identity verification must be one of the major pain points for users (who have to reverify their images if initially rejected). **Automated prompts** when the picture is being taken like *'face not visible', 'not enough light', 'face is too far', 'card out of focus'* etc. may help genuine users to click proper pictures or rectify some shortcomings of the image that may not be apparent to the user

Taking this a step forward, the correct picture is automatically taken when the image settings is detected to be ok after the user places his/her face or card/document in the shape rather than expect them to click the shutter button.

**Session level Conversion Rate , MIR, NPS scores** for the Verify Identity stage could be expected to improve immediately. **Number of Sessions per Verification, Number of Rejections per Verification, Entry Rate** for the Verify Identity stage could be expected to decline indicating that user experienced is helped by such prompts.

**Dwell Time** for Verify Identity stage may increase as prompts may cause users to try and update their current picture configurations, though the inconvenience should be compensated by not having to redo the verification process

This should also improve **Conversion Rate (user level)** but it may take time to notice the uplift. **Fraud detection error** rates should remain unaffected but it's always a good idea to keep a close eye on the metric post introduction of any new feature.

## Hypothesis 2: Auto-detection of User Details from id (either using the image scanned or using the NFC)

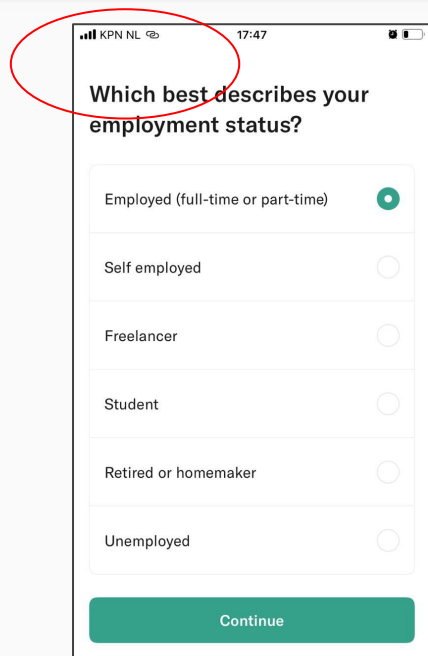
- Auto-Detection of user details from Id scan (or NFC read):

Even though this is before the email-id confirmation, manually entering all the user details when the user eventually is going to have to provide his/her id (which has all such details) seems redundant. Ideally, one could expect that information **automatically extracted** from the **id scan (or using the NFC in the id)** to be pre populated which can then be verified by the user. **Quick checks** to ensure the validity of the id (like a valid expiration date) could also happen in real time. This ensures minimal intervention from the user, minimal chances of user mistakes, and improper documents rejected at real time before phone numbers are even registered.

**Dwell Time, NPS scores** for the initial stage should reduce as a result of such Auto detection. Average **Number of Sessions per Verification, Number of Rejections per Verification** should also reduce since simple rejections because of mismatch of ids and details submitted would be minimised

**Conversion Rate (user level)** could improve because users are not going to be turned away due to petty errors while their emails and phone numbers get registered. However the increase would only be marginal at best

## Hypothesis 3: Navigation buttons doesn't show



The screenshot shows a mobile app interface for a KYC questionnaire. At the top, the status bar displays 'KPN NL' and the time '17:47'. Below this, the question 'Which best describes your employment status?' is displayed. There are six radio button options: 'Employed (full-time or part-time)' (selected), 'Self employed', 'Freelancer', 'Student', 'Retired or homemaker', and 'Unemployed'. At the bottom, there is a green 'Continue' button. A red circle highlights the 'KPN NL' text in the status bar, indicating the back button.

- Back (or other Navigation) buttons don't show during for KYC questionnaire:

Even though the back button is present, it doesn't show up against the white background and this ensues significant confusion for the user if they decide to change or revise their earlier entries

Drop Off rates may reduce if these navigational elements show up. **Dwell Time** for the Answer Questions Stage will increase but will improve **NPS scores**

May have a Long term impact on the **Fraud detection error (False Positive) rate** since genuine users now have a greater likelihood of providing more accurate information

# Thanks!

-Subhra

