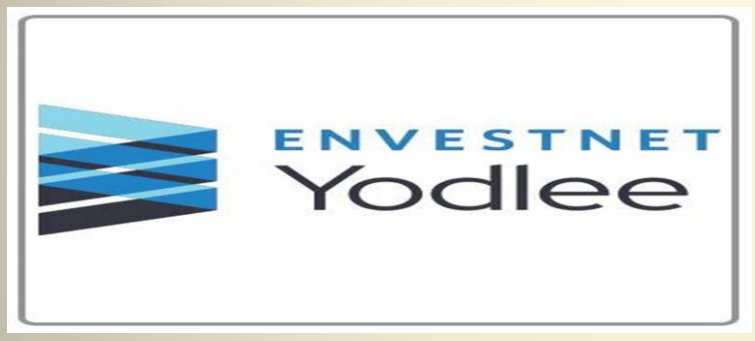
SIX MONTH INTERNSHIP THESIS REPORT

NAME-ARUNAVA DHAR

Roll No- CRS1904

Mtech in Cryptology & Security



ARUNAVA DHAR

BULK CATEGORIZATION ANALYZER

Primary Supervisor- **Minu Catherine Susainathan**

Secondary Supervisor- **Anisur Rahman**

THESIS REPORT

I am currently working as an intern in **ENVESTNET YODLEE** as a **PROJECT TRAINEE. I have** been appointed in QA team and I am working under my primary supervisor **Minu Catherine Susainathan** and my secondary supervisor is **Anisur Rahman Sir.**

Problem Statement:

In order to maintain our product here each and every day automation tests are being executed by our team. These tests are executed against multiple environments to check the integrity of the software. So when executed against these environments when the underlying product component is down or slow the failures in these automation executions becomes huge and tedious.
So my goal is to identify the reason behind these bulk failures and provide a suitable solution to it.

- The automation execution generally takes hours to get completed
- If we get to know before hand that the bulk failures are due to environment issues, there is no point of continuing the suite execution
- Blockage of resources
- Wastage of time in running the execution and analyzing it further

Significance of the problem:

While testing these under various environments lot of human effort is required and thus in turn lot of resources is also required in automation execution and triaging of these failures

So if the failure is in bulk that is more than 40% of the test cases are failing then the cost to analyse and rectify these failures will be very high and also the environment stability will be fixed very late.

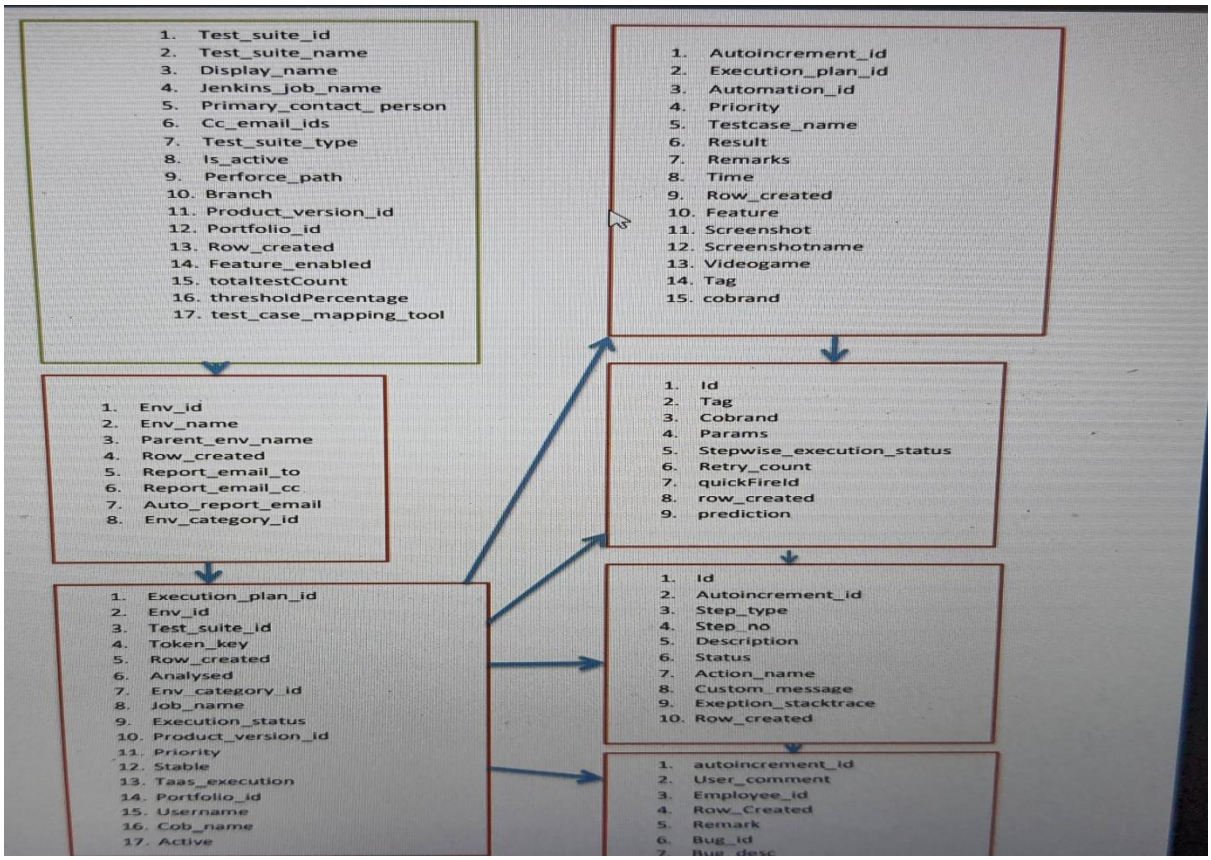
So if the identification of this bulk failure can be made then huge amount of resources can be saved.

Approach to the problem:

First we need to identify all the failure exceptions and their corresponding category. We need to try to create a Machine Learning model that will take those data as input and process the data and in future predict the reason for failures. During automation execution at predefined checkpoints the failed percentage will be validated, post the threshold is crossed, the data will be given to the model and the corresponding categories will be returned. And based on the tickets raised the deployment team will resolve the issue.

Progress made to solve the problem:

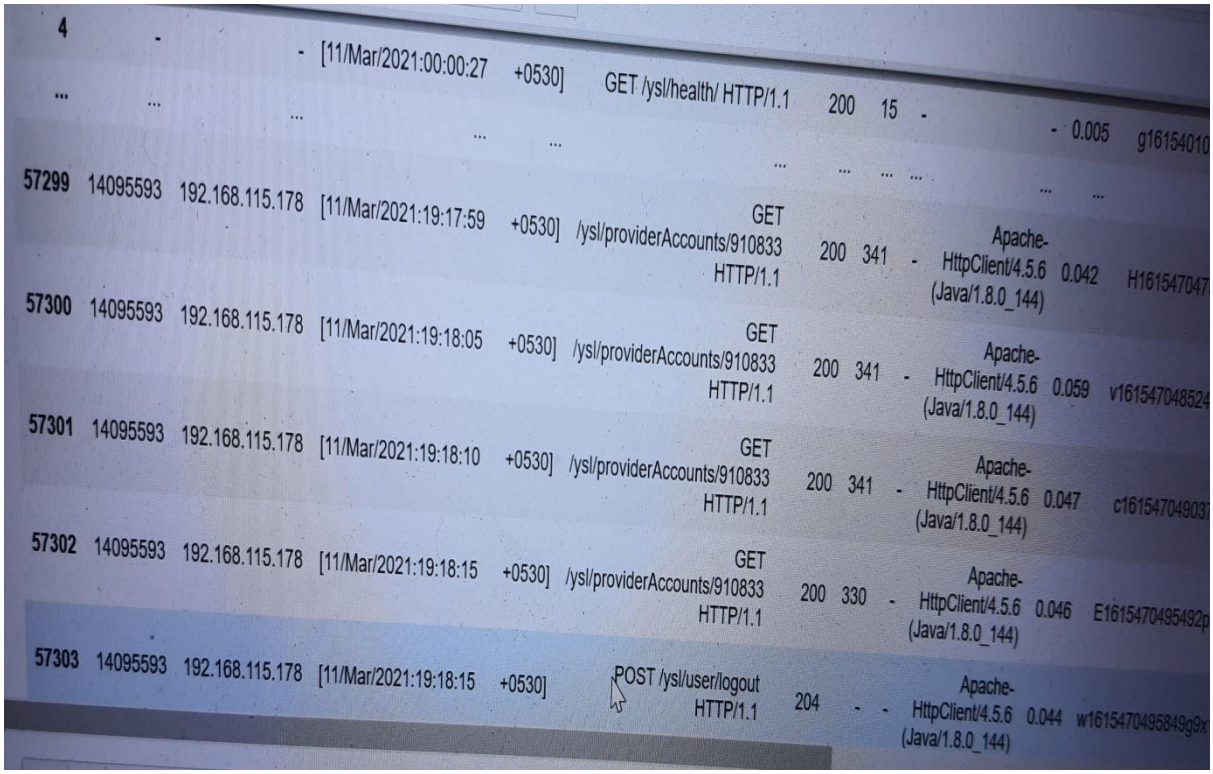
Firstly I was given a first hand demonstration of the data base that I had to handle and was asked to make a Relational database diagram. The database was on the Reporting dashboard and I was asked to make a Reporting Dashboard relationship Model to get a better understanding of the data that I would be working on with at a later stage.



Now everyday lots of automation tests are being performed and for those automation tests employees need to send url requests for those particular softwares. Firstly I was given a demo on how each API request for a particular software was passed using the app POSTMAN. Over there for a particular product of the company different APIs were hit and I was asked to download those logs of each API hit and study those logs and search for some pattern.

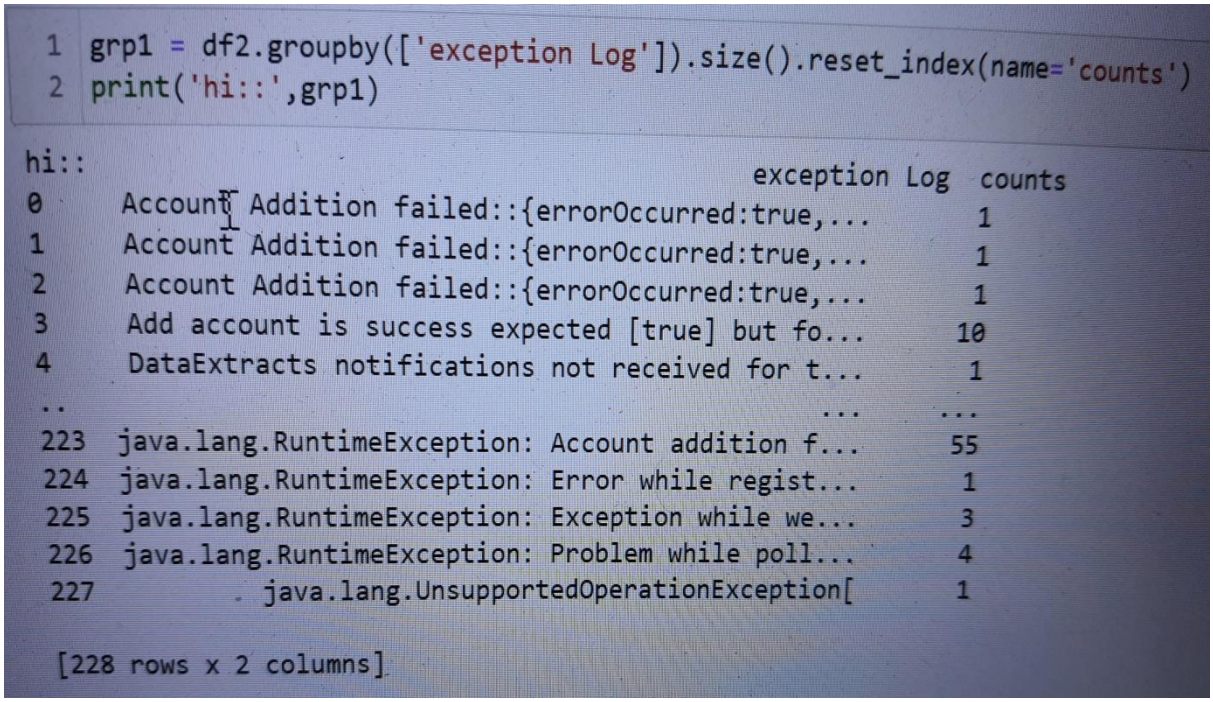
In our company software User registration and various other jobs is done for various banks. So I was asked to study a particular test case comprising of different APIs for different users and was asked to find a pattern between them i.e how the log request looked for different users.

I found that the difference in the log took place in what is called the **member ID.** So I was asked to write a program in order to group the member id and print the logs according to the member ID.



Our company holds lots of product like Money Centre , fastlink etc and everyday lots of automation tests are being performed for the betterment of these softwares. So while performing those automations lots of exceptions happens. So I was asked to group all these exceptions according to the nature of the exceptions. While performing tests on the different APIs of these apps lots of bulk failure happens. And I am given to study the reason of these bulk failures.

So I had been given data of the previous months of these softwares. I had to perform a detailed study of these exceptions happening. Then I had to perform pre-processing of the data where I had to trim down the exceptions and slit out the important part of those exceptions. After that I had to group the exceptions using 'Groupby()' method.



```
1 grp1 = df2.groupby(['exception Log']).size().reset_index(name='counts')
2 print('hi::',grp1)
```

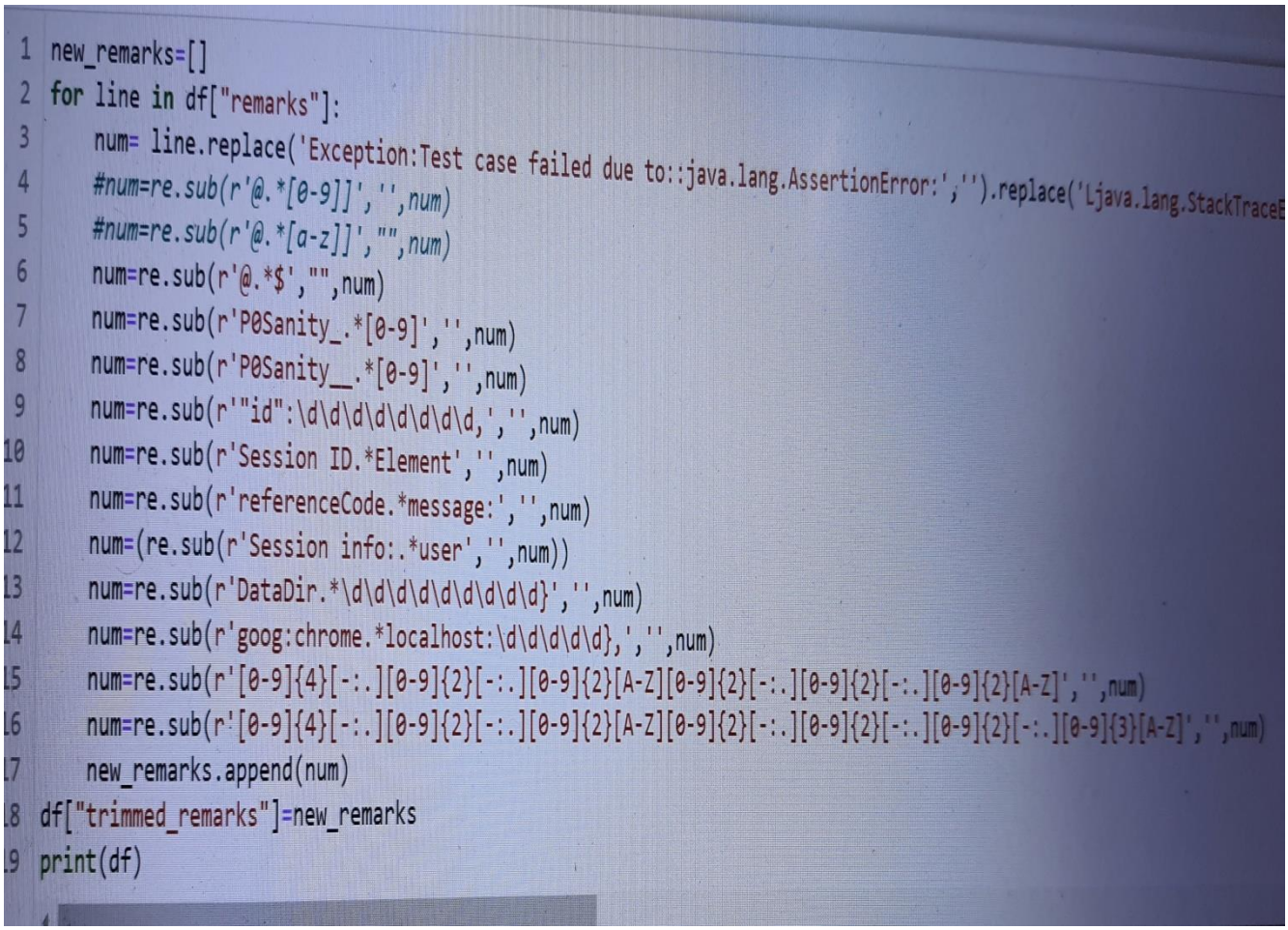
```
hi::
                                                exception Log  counts
0     Account Addition failed::{errorOccurred:true,...       1
1     Account Addition failed::{errorOccurred:true,...       1
2     Account Addition failed::{errorOccurred:true,...       1
3     Add account is success expected [true] but fo...      10
4     DataExtracts notifications not received for t...       1
..                                                  ...     ...
223  java.lang.RuntimeException: Account addition f...      55
224  java.lang.RuntimeException: Error while regist...       1
225  java.lang.RuntimeException: Exception while we...       3
226  java.lang.RuntimeException: Problem while poll...       4
227          java.lang.UnsupportedOperationException[       1

[228 rows x 2 columns]
```

After this with lots of preprocssing in hand the data had to be cleaned thoroughly for using it in an NLP model.

Pre Processing of data

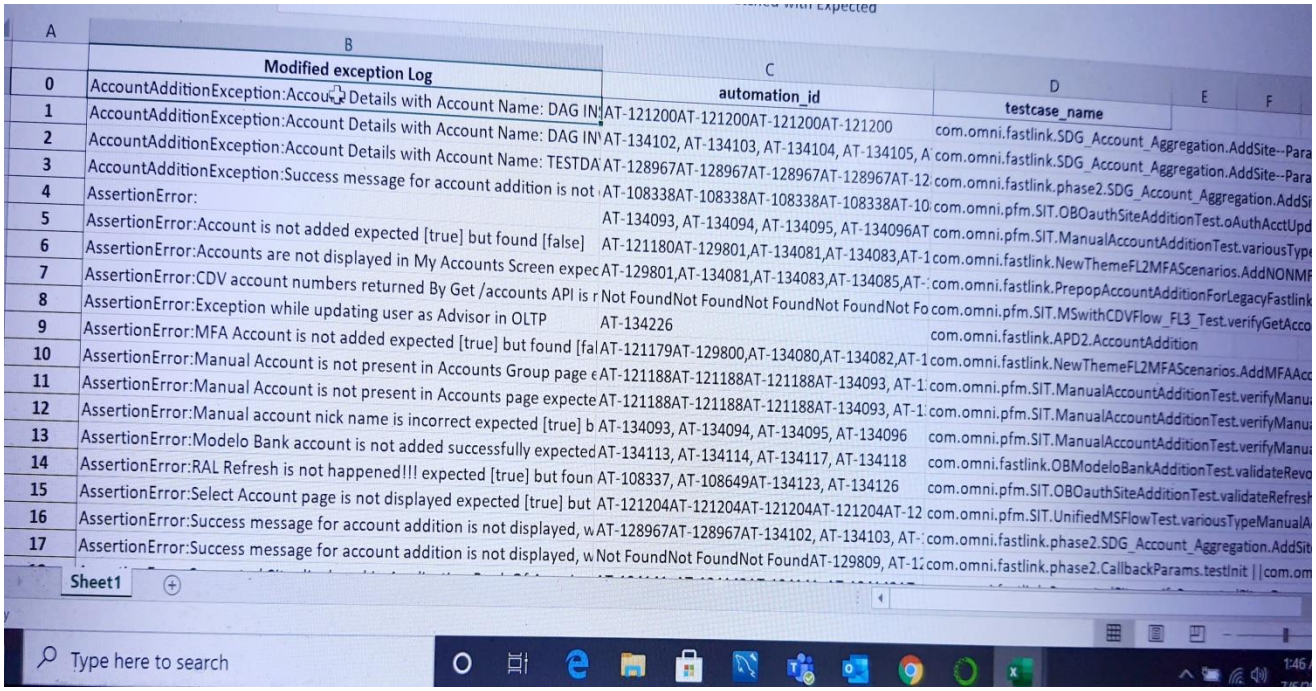
The data that was given by the company was complete raw data with lots of anomalies and noise in the data. The data had to be thoroughly pre-processed before it could be actually used as a training dataset. The groups that was made using the groupby() method had to be cleaned thoroughly removing the noise so that a higher level subgrouping can be done. The data was cleaned programmatically.

The steps of cleaning are given in the picture below:



```
new_remarks=[]
for line in df["remarks"]:
    num= line.replace('Exception:Test case failed due to::java.lang.AssertionError:','').replace('Ljava.lang.StackTraceE
    #num=re.sub(r'@.*[0-9]]','',num)
    #num=re.sub(r'@.*[a-z]]',"",num)
    num=re.sub(r'@.*$',"",num)
    num=re.sub(r'P0Sanity_.*[0-9]','',num)
    num=re.sub(r'P0Sanity__.*[0-9]','',num)
    num=re.sub(r'"id":\d\d\d\d\d\d\d\d,','',num)
    num=re.sub(r'Session ID.*Element','',num)
    num=re.sub(r'referenceCode.*message:','',num)
    num=(re.sub(r'Session info:.*user','',num))
    num=re.sub(r'DataDir.*\d\d\d\d\d\d\d\d\d}','',num)
    num=re.sub(r'goog:chrome.*localhost:\d\d\d\d\d},','',num)
    num=re.sub(r'[0-9]{4}[-:.][0-9]{2}[-:.][0-9]{2}[A-Z][0-9]{2}[-:.][0-9]{2}[-:.][0-9]{2}[A-Z]','',num)
    num=re.sub(r'[0-9]{4}[-:.][0-9]{2}[-:.][0-9]{2}[A-Z][0-9]{2}[-:.][0-9]{2}[-:.][0-9]{2}[-:.][0-9]{3}[A-Z]','',num)
    new_remarks.append(num)
df["trimmed_remarks"]=new_remarks
print(df)
```

The data comprised of different test suites that were being performed of the various products. These test suites comprises a unique **automation id & auto increment id &** other columns such as **test case details, test case name, exception log & results.** After performing the cleaning of the data a new column was added which was the **Modified exception logs. T**hat new data had to be stored in a csv file file that could be used later for other purposes.

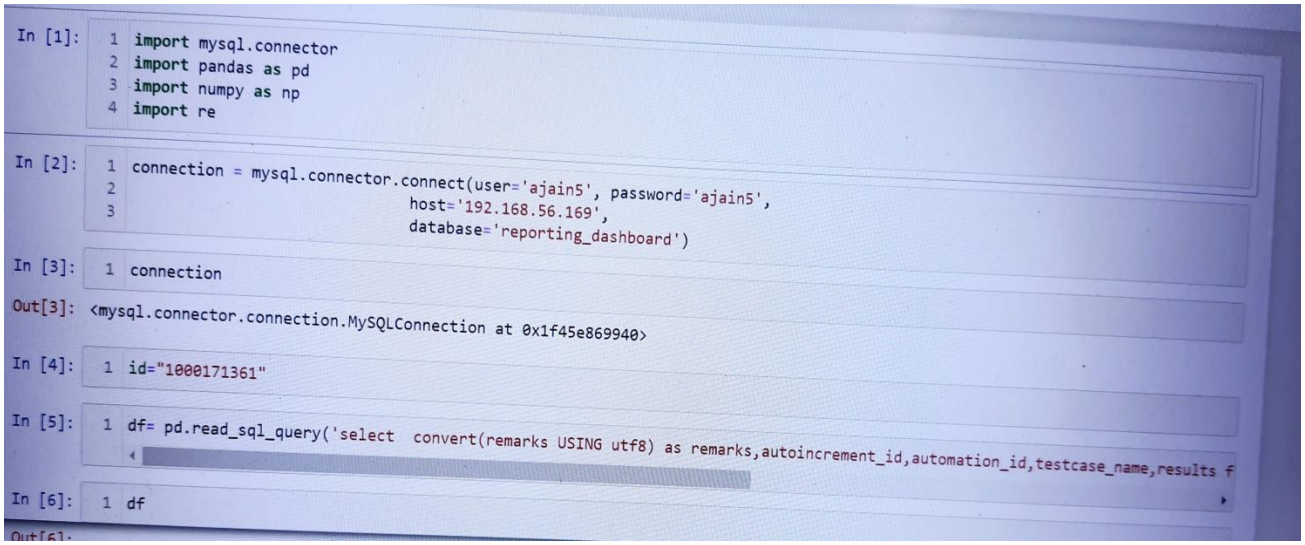


	A	B	C	D	E	F
		Modified exception Log	automation_id	testcase_name		
0		AccountAdditionException:Accou Details with Account Name: DAG IN	AT-121200AT-121200AT-121200AT-121200	com.omni.fastlink.SDG_Account_Aggregation.AddSite--Para		
1		AccountAdditionException:Account Details with Account Name: DAG IN	AT-134102, AT-134103, AT-134104, AT-134105, A	com.omni.fastlink.SDG_Account_Aggregation.AddSite--Para		
2		AccountAdditionException:Account Details with Account Name: TESTDA	AT-128967AT-128967AT-128967AT-128967AT-12	com.omni.fastlink.phase2.SDG_Account_Aggregation.AddSi		
3		AccountAdditionException:Success message for account addition is not	AT-108338AT-108338AT-108338AT-108338AT-10	com.omni.pfm.SIT.OBOauthSiteAdditionTest.oAuthAcctUpd		
4		AssertionError:	AT-134093, AT-134094, AT-134095, AT-134096AT	com.omni.pfm.SIT.ManualAccountAdditionTest.variousType		
5		AssertionError:Account is not added expected [true] but found [false]	AT-121180AT-129801,AT-134081,AT-134083,AT-1	com.omni.fastlink.NewThemeFL2MFAScenarios.AddNONMF		
6		AssertionError:Accounts are not displayed in My Accounts Screen expec	AT-129801,AT-134081,AT-134083,AT-134085,AT-	com.omni.fastlink.PrepopAccountAdditionForLegacyFastlink		
7		AssertionError:CDV account numbers returned By Get /accounts API is r	Not FoundNot FoundNot FoundNot FoundNot Fo	com.omni.pfm.SIT.MSwithCDVFlow_FL3_Test.verifyGetAcco		
8		AssertionError:Exception while updating user as Advisor in OLTP	AT-134226	com.omni.fastlink.APD2.AccountAddition		
9		AssertionError:MFA Account is not added expected [true] but found [fal	AT-121179AT-129800,AT-134080,AT-134082,AT-1	com.omni.fastlink.NewThemeFL2MFAScenarios.AddMFAAcc		
10		AssertionError:Manual Account is not present in Accounts Group page e	AT-121188AT-121188AT-121188AT-134093, AT-1	com.omni.pfm.SIT.ManualAccountAdditionTest.verifyManu		
11		AssertionError:Manual Account is not present in Accounts page expecte	AT-121188AT-121188AT-121188AT-134093, AT-1	com.omni.pfm.SIT.ManualAccountAdditionTest.verifyManu		
12		AssertionError:Manual account nick name is incorrect expected [true] b	AT-134093, AT-134094, AT-134095, AT-134096	com.omni.pfm.SIT.ManualAccountAdditionTest.verifyManu		
13		AssertionError:Modelo Bank account is not added successfully expected	AT-134113, AT-134114, AT-134117, AT-134118	com.omni.fastlink.OBModeloBankAdditionTest.validateRevo		
14		AssertionError:RAL Refresh is not happened!!! expected [true] but foun	AT-108337, AT-108649AT-134123, AT-134126	com.omni.pfm.SIT.OBOauthSiteAdditionTest.validateRefresh		
15		AssertionError:Select Account page is not displayed expected [true] but	AT-121204AT-121204AT-121204AT-121204AT-12	com.omni.pfm.SIT.UnifiedMSFlowTest.variousTypeManualA		
16		AssertionError:Success message for account addition is not displayed, w	AT-128967AT-128967AT-134102, AT-134103, AT-	com.omni.fastlink.phase2.SDG_Account_Aggregation.AddSit		
17		AssertionError:Success message for account addition is not displayed, w	Not FoundNot FoundNot FoundAT-129809, AT-1	com.omni.fastlink.phase2.CallbackParams.testInit ||com.om		

Sheet1

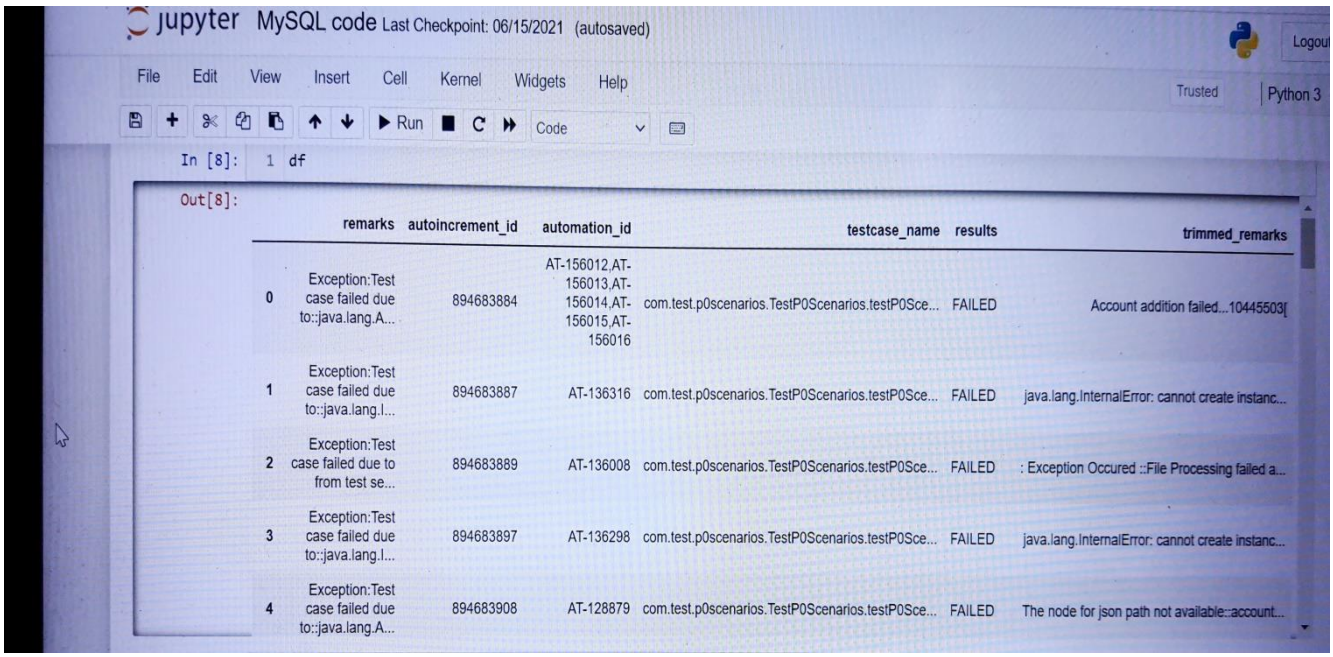
Type here to search

Till now everything that had been done was done by downloading the data from the server and applying everything to it. But in realitity when this has to be to be done, I have to directly take the data from the SQL server and perform operations on it. For that I was told to create a connection from the SQL server to PYTHON in order to fetch the data from the server and apply the necessary pre-processing to it.



```
In [1]:
import mysql.connector
import pandas as pd
import numpy as np
import re

In [2]:
connection = mysql.connector.connect(user='ajain5', password='ajain5',
                                     host='192.168.56.169',
                                     database='reporting_dashboard')

In [3]:
connection

Out[3]: <mysql.connector.connection.MySQLConnection at 0x1f45e869940>

In [4]:
id="1000171361"

In [5]:
df= pd.read_sql_query('select  convert(remarks USING utf8) as remarks,autoincrement_id,automation_id,testcase_name,results f

In [6]:
df
```

The pre-processed data was stored in a new column called **‘trimmed remarks’.**



jupyter MySQL code Last Checkpoint: 06/15/2021 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

▶ Run Code

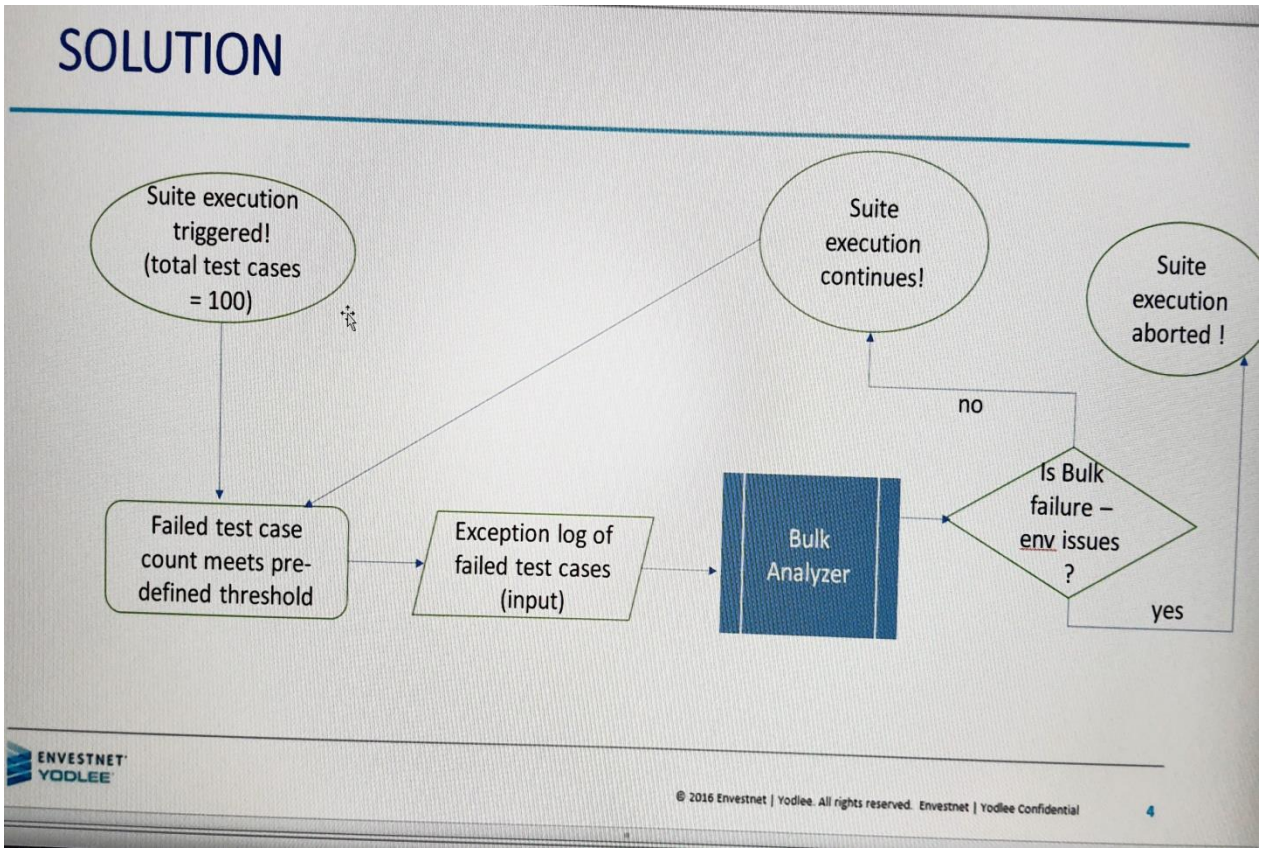
In [8]: 1 df

Out[8]:

	remarks	autoincrement_id	automation_id	testcase_name	results	trimmed_remarks
0	Exception:Test case failed due to::java.lang.A...	894683884	AT-156012,AT-156013,AT-156014,AT-156015,AT-156016	com.test.p0scenarios.TestP0Scenarios.testP0Sce...	FAILED	Account addition failed...10445503[
1	Exception:Test case failed due to::java.lang.I...	894683887	AT-136316	com.test.p0scenarios.TestP0Scenarios.testP0Sce...	FAILED	java.lang.InternalError: cannot create instanc...
2	Exception:Test case failed due to from test se...	894683889	AT-136008	com.test.p0scenarios.TestP0Scenarios.testP0Sce...	FAILED	: Exception Occured ::File Processing failed a...
3	Exception:Test case failed due to::java.lang.I...	894683897	AT-136298	com.test.p0scenarios.TestP0Scenarios.testP0Sce...	FAILED	java.lang.InternalError: cannot create instanc...
4	Exception:Test case failed due to::java.lang.A...	894683908	AT-128879	com.test.p0scenarios.TestP0Scenarios.testP0Sce...	FAILED	The node for json path not available::account...

Solution to the problem in Hand

- Bulk Analyzer will take the exception log of the failed test cases as input.
- It will analyze the exceptions and make a decision if the bulk failures are due to environment issues or not
- Based on the decision, it will direct for abortion/continuation of the suite execution.



RULES FOR BULK CATEGORIZATION IMPLEMENTATION

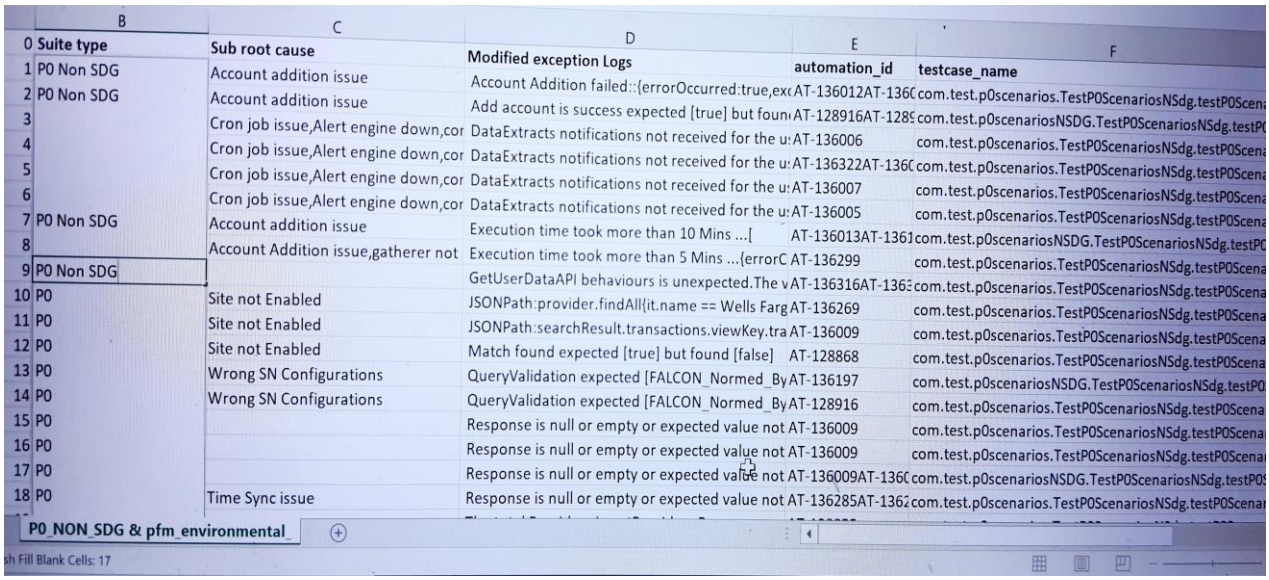
- A pre-defined value will be declared, say 25%. It means, only after 25% of the execution gets completed, bulk analyzer will kick in.
- A threshold failure% will be declared. So, after the failure% reaches the threshold, bulk analyzer will get triggered.

NOTE: New rules will be added later and the existing rules can be enhanced henceforth.

TRAINING DATA SET PREPARATION

1. Dump taken of the failure test case history for the targeted suites
2. Noise removal from the exception logs of the failed test cases
3. Manual analysis of the exceptions and finding out the sub root cause analysis
4. Data set prepared off the exception log and its sub root cause analysis

A Dictionary type data structure is being created using the Modified exception logs. The sub root caused is manually created by analysing the exception logs. Later in the model the modified exception logs will be put a s training data set which with the help of NLP will bring out the sub root causes which will in turn help in analysing the bulk failure reason

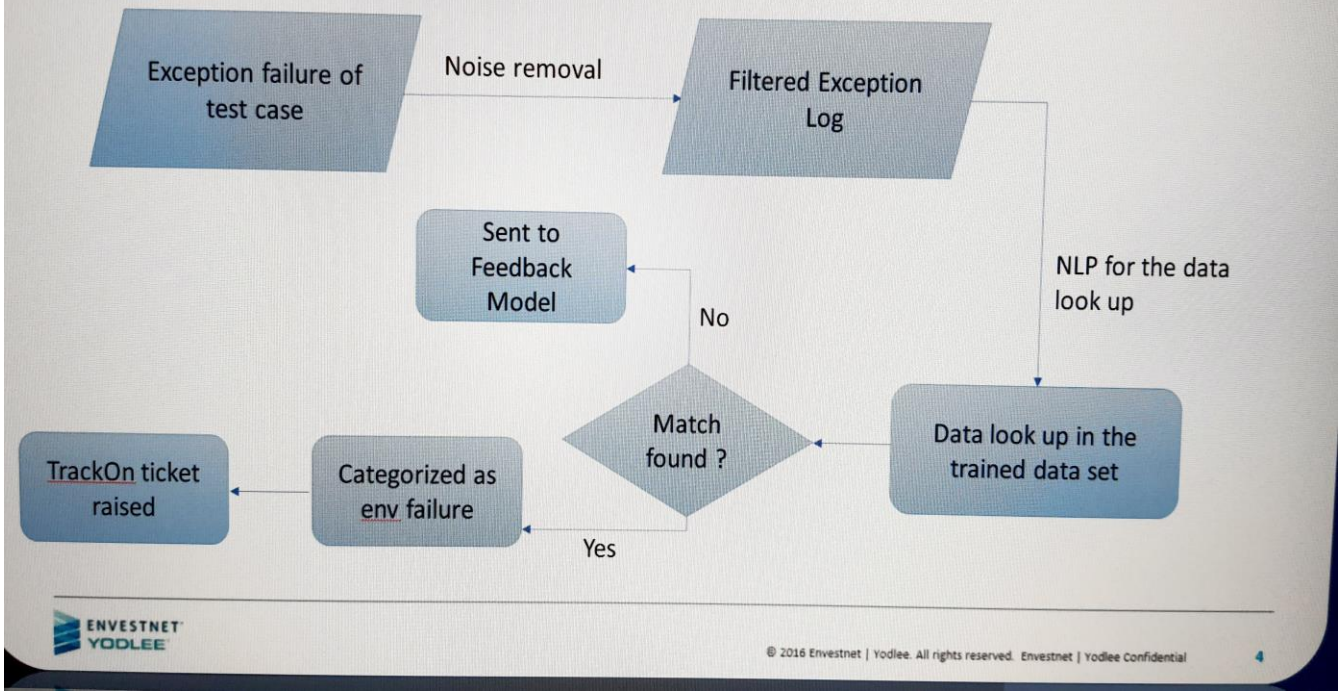


	B	C	D	E	F
0	Suite type	Sub root cause	Modified exception Logs	automation_id	testcase_name
1	P0 Non SDG	Account addition issue	Account Addition failed::{errorOccurred:true,exc	AT-136012AT-136(	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
2	P0 Non SDG	Account addition issue	Add account is success expected [true] but foun	AT-128916AT-128	com.test.p0scenariosNSDG.TestP0ScenariosNSdg.testP(
3		Cron job issue,Alert engine down,cor	DataExtracts notifications not received for the u	AT-136006	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
4		Cron job issue,Alert engine down,cor	DataExtracts notifications not received for the u	AT-136322AT-136(	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
5		Cron job issue,Alert engine down,cor	DataExtracts notifications not received for the u	AT-136007	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
6		Cron job issue,Alert engine down,cor	DataExtracts notifications not received for the u	AT-136005	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
7	P0 Non SDG	Account addition issue	Execution time took more than 10 Mins ...[	AT-136013AT-1361	com.test.p0scenariosNSDG.TestP0ScenariosNSdg.testP(
8		Account Addition issue,gatherer not	Execution time took more than 5 Mins ...{errorC	AT-136299	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
9	P0 Non SDG		GetUserDataAPI behaviours is unexpected.The v	AT-136316AT-136	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
10	P0	Site not Enabled	JSONPath:provider.findAll{it.name == Wells Farg	AT-136269	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
11	P0	Site not Enabled	JSONPath:searchResult.transactions.viewKey.tra	AT-136009	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
12	P0	Site not Enabled	Match found expected [true] but found [false]	AT-128868	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
13	P0	Wrong SN Configurations	QueryValidation expected [FALCON_Normed_By	AT-136197	com.test.p0scenariosNSDG.TestP0ScenariosNSdg.testP0
14	P0	Wrong SN Configurations	QueryValidation expected [FALCON_Normed_By	AT-128916	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
15	P0		Response is null or empty or expected value not	AT-136009	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
16	P0		Response is null or empty or expected value not	AT-136009	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena
17	P0		Response is null or empty or expected value not	AT-136009AT-136(	com.test.p0scenariosNSDG.TestP0ScenariosNSdg.testP0
18	P0	Time Sync issue	Response is null or empty or expected value not	AT-136285AT-136	com.test.p0scenarios.TestP0ScenariosNSdg.testP0Scena

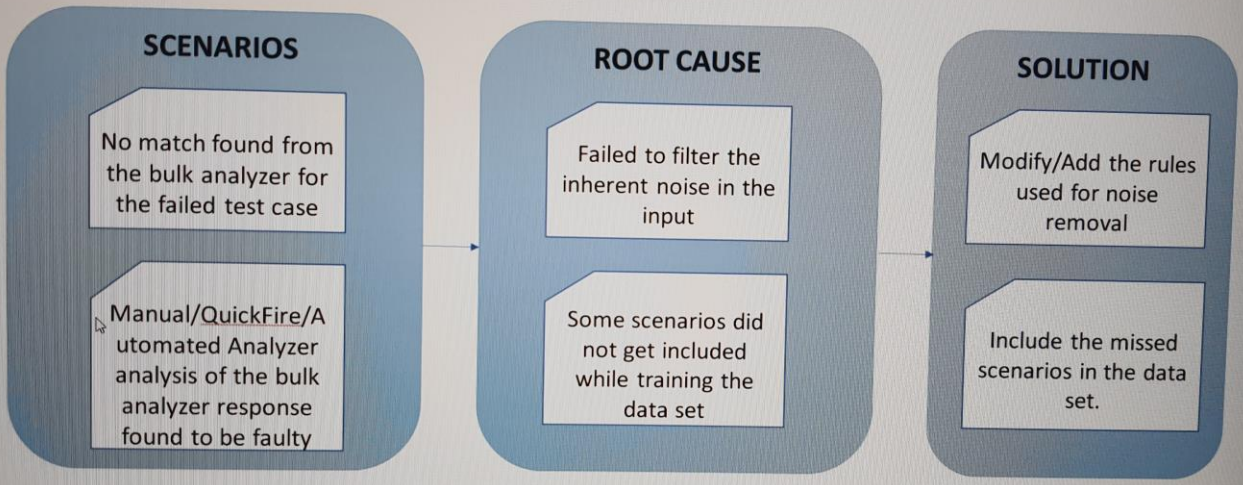
P0_NON_SDG & pfm_environmental

Fill Blank Cells: 17

PROCESS FLOW



FEEDBACK MODEL

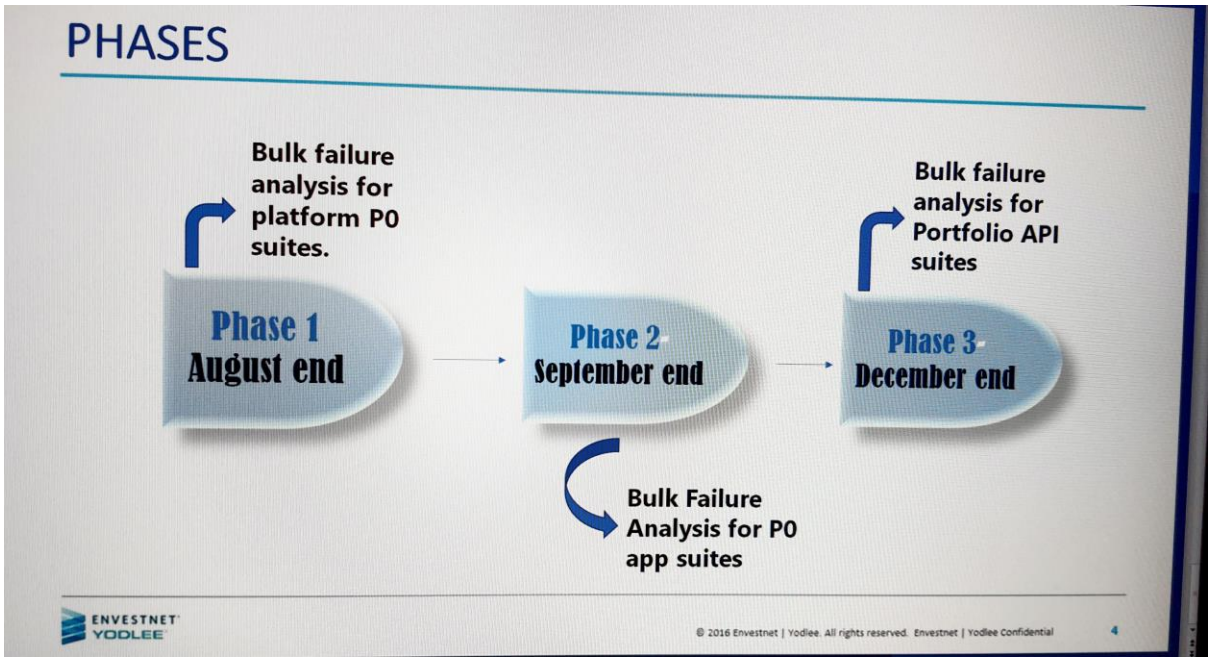


Upcoming Work

The Data set has been prepared for 2 test suites namely PFM and P0 NON SDG. More data sets for other test suites need to be prepared so that the dictionary can have vast range of input data to train and for the model to predict.

The NLP model will be created in the upcoming months and the data will be tested according to that.

The project contains 3 phases as follows:



Phase 1 is almost completed & will be put up in the upcoming month.

THANKS & REGARDS

I want to thank Indian Statistical Institute for providing me this great opportunity to work in a prestigious company as Envestnet Yodlee.

A hearty Thank you to my Primary Supervisor **MINU CATHERINE SUSAINATHAN.** She has been a constant support throughout my whole internship.

I have been offered a FULL TIME EMPLOYEE (FTE) in this company and I am pretty much excited & looking forward to working under her guidance. She has been a great mentor to me.

I also want to thank the members of the QA team specially **DIKHSHA AGARWAL** who is also a part of this project and has helped me a lot to get adapted & also with this project**.**

A special thanks to Anisur Sir who has helped me lot in this journey.

THE END