Financial Crime Analysis in Power BI

# Revision Number: 1.0

Last date of revision: 18/05/2022

Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Date Issued | Version | Description | Author |
| 18/05/2022 | 1 | Initial HLD — V1.0 | Prateek Jha |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Contents**

[Document Version Control 2](#_TOC_250021)

[Abstract 4](#_TOC_250020)

1. [Introduction 5](#_TOC_250019)
   1. [Why this High-Level Design Document? 5](#_TOC_250018)
   2. [Scope 5](#_TOC_250017)
   3. [Definitions 5](#_TOC_250016)
2. [General Description 6](#_TOC_250015)
   1. [Product Perspective 6](#_TOC_250014)
   2. [Problem statement 6](#_TOC_250013)
   3. [PROPOSED SOLUTION 6](#_TOC_250012)
   4. [FURTHER IMPROVEMENTS 6](#_TOC_250011)
   5. [Technical Requirements 6](#_TOC_250010)
   6. [Data Requirements 7](#_TOC_250009)
   7. [Tools used 8](#_TOC_250008)
      1. Hardware Requirements 8
      2. ROS(Robotic Operating System) 9
   8. [Constraints 9](#_TOC_250007)
   9. [Assumptions 9](#_TOC_250006)
3. [Design Details 10](#_TOC_250005)
   1. [Process Flow 10](#_TOC_250004)
      1. [Model Training and Evaluation 10](#_TOC_250003)
      2. [Deployment Process 11](#_TOC_250002)
   2. [Event log 11](#_TOC_250001)
   3. [Error Handling 11](#_TOC_250000)
   4. Performance 12
   5. Reusability 12
   6. Application Compatibility 12
   7. Resource Utilization 12
   8. Deployment 12
4. Dashboards 13
   1. KPls (Key Performance lndicators) 13
5. Conclusion 14

# Abstract

Financial institutions around the world are turning to data science to combat crime and manage compliance due to the changing nature of crime and a quickly expanding regulatory landscape.

The global financial crisis of 2008 altered the course of history. It had an impact not only on the financial industry, but also on other industries and enterprises around the world. The crisis exposed ineffective policies that resulted in severe fractures that threatened to bring the global financial system to its knees.

Technological advancements, and new capabilities to understand enormous volumes of data can help to analyse and formulate the best approach to identify flaws and appropriate interventions techniques to reduce financial crime.

# Introduction

## **Why this High-Level Design Document:**

The main goal of this high level design (HLD) document is to detect all the inconsistencies in the design prior to code implementation and it can be used as a reference manual in case we need to understand the working of any manual for any of the modules.

This HLD will contain the following information needed:

* + - Present all of the design aspects and define them in detail
    - Describe the user interface being implemented
    - Description of user interfaces
    - Description of performance requirements
    - Project design as a whole unit
    - List and describe the non-functional attributes like:
      * + Reliability
        + Maintainability
        + Portability
        + Reusability
        + Application compatibility
        + Resource utilization
        + Serviceability

## Scope

This HLD documentation covers the overall architecture of the system, such as the database usage, application’s flow layer by layer both from UI and backend perspective and technology architecture. This HLD will contain non-technical to slightly technical terms related to the work but we will explain all of this here completely from scratch without any problem.

## Definitions

*Term Description*

|  |  |
| --- | --- |
| *PBI*  *DB*  *ETL*  *CSV* | Microsoft Power BI |
| DB is the short form for database or collection of all the information monitored by this system |
| Used to denote the term Extract Transform and Load |
| Source files containing data to be put in MySQL database |

# General Description

### **Problem statement**

We are trying to give the users a self service BI report that will give the users a brief idea as to help them focus and try to find the reasons and numbers that support the details about fraudulent transactions that are happening based on the data provided for different accounts and one or multiple transactions that happen and alerts recorded for certain transactions

It is very difficult to know about the overall idea just by looking at the raw file data in best possible manner but everyone understands data better in a visually appealing way. So we are designing a report in Power BI which would resolve all those issues in understanding data in raw form as it is generated.

### **PROPOSED SOLUTION**

We are proposing a solution where users with no or little background knowledge of self service report such as Power BI, one can simply come to this report that we’ve opened for our users, enter some filters on desired values based on the preferences from available list of options can get the most accurate results for available data with very less/no deviation.

The proposed solution here is a PBI report which is hosted in Power BI service which will capture all data changes from the source files generated for the same, perform an ETL using python and query editor steps in PBI itself to clean the data as much we can and using the logic appropriately for different visuals for representing different needs.

### **FURTHER IMPROVEMENTS**

Of course, this report will need to be refreshed with more data so a new file will be needed everytime with updated data which will be processed via python script to be dumped in MySQL DB and then finally to be refreshed in PBI service which is on cloud.

## **Technical Requirements**

* + - Data requirement completely depend on our problem statement.
    - We need data in a relational table format containing various columns/features where we will have one or multiple columns which will be used for accurate display of information.
    - We will need some columns as independent columns such as Account\_ID, Account\_Type, IS\_FRAUD(flag column with Boolean values), Alert\_ID, TX\_ID(TransactionID), TX\_AMT(TransactionAmount) and many others across the 3 different tables
    - The source of all this data will be 3 CSV files and after collection of entire data, it will dumped into a MySQL DB after some cleaning is done via python script.
    - As far as technical knowledge is concerned, we need to be fundamentally cleared up on basics of some DAX functions in PBI for calculated columns and measures, pandas and python data manipulation using lists
    - Working knowledge of MySQL DB as this is the main database which we will be using for data related activities.

## **Data Requirements**

* + - Data requirement completely depend on our problem statement.
    - We data in a relational table format containing various columns/features where we will have one or multiple columns which will be used for prediction and there will be only one dependent feature/column which will be used for prediction.
    - The source of all these files will be a CSV file and after collection of entire data, it will dumped into a MySQL DB and from there, we will fetch it for cleaning and import in PBI accordingly needed per purpose.

## **Tools used**

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn are used to build the entire backend part for machine learning modelling part.



* + - IDE to be used for development for Python script for ETL purpose is Jupyter notebook or any IDE that support python scripting and we can schedule to run this python script at regular intervals and once the data is loaded is to be used in Power BI for display.
    - MySQL DB is used for storage and retrieval from source database.
    - GitHub is used as version control system and contains the repository for this one.
    - Microsoft Power Bi for April 2022 release is currently used for the development of the report but can be used for changes in any upcoming versions of the release for this product.
* ~~- - - •~~ High Level Design 

## **Constraints**

The report created should be user friendly and should provide the user with correct details for which user is using the platform for with best accurate data to be shown in visual format.

**- 1**High Level Design (HLD) 

# Design Details

## **Process Flow**

For our Fraudulent transactions analysis report to work properly, we will use a python script to fetch data from csv files and rest of steps and explained in detail below

Proposed methodology

