DATA QUALITY DASHBOARD

INTRODUCTION

THE BACKGROUND

Polen Capital is a fully integrated global investment management firm, whose business model revolves around acquiring, analysing, and distributing high quality data. These data elements with unique uses throughout the organization, and multiple downstream dependencies, comes with a high severity associated with errors and failures. As part of The Capstone Project at RIT, we were asked to build a Dashboard that would help the Business Stewards identify and act on data quality issues. The guiding principles of data at Polen Capital are as follows:

1. Data is a highly valued enterprise asset
2. Critical data will be managed by domains rather than systems and be made available through a source of truth (trusted, primary sources of data). This includes data generated by a third party.
3. Critical data should be clearly defined by domain and element (e.g., origin, meaning and quality) to ensure its properly accessed and consumed
4. Data Quality issues should be communicated to the domain owners to ensure they are corrected at the system of record and source of truth.
5. Data Quality should be run at the source of truth and post transformations to ensure downstream systems receive quality data.

THE OBJECTIVE

The primary users of our dashboard are the cross-functional Business Stewards (SMEs, and Custodians) who act as conduits between the Data Providers and the Consumers. Based on the brief and the data provided we identified three key requirements that can help the Business Stewards identify Data Quality issues:

1. A feature/ tool to explain the overall performance of the data across in terms of FAIL & PASS records.
2. A feature/tool to see the Location, Magnitude, and the Trend of the errors.
3. A feature/tool to do a Root-Cause Analysis and derive insights.

DESCRIPTION OF THE DATA

The data so far has come in 6 batches of Excel files, ranging from 29th May 2023 till 17th July 2023, and with the following columns:

Column	Type	Description
AIM_ID	Integer	Categorical variable containing unique IDs, representing the date the data was run. This corresponds to the ‘Check Timestamp’ (Column G).
CHECK_ID	Integer	Categorical. Unique IDs represent the data quality check.
CHECK_TYPE	String	Contains only one type: CDE (Critical Data Element).
CHECK_DATASET	String	Type of Datasets where the data is checked
CHECK_COLUMN	String	The field that is being checked
CHECK_MESSAGE	String	Represents what is being checked within the column.
CHECK_TIMESTAMP	Datetime	Timestamp on which check occurred
CHECK_SQL	String	SQL query used for checking
CHECK_SYSTEM	String	3 unique Systems: ARW, DATAHUB, PCDW
PASS_COUNT	Integer	Number of records passed the check
PASS_PCT	Double	Percentage of passed records
FAIL_COUNT	Integer	Number of records failed the check
FAIL_PCT	Double	Percentage of failed records
RISK CATEGORY*	String	Categorical. Divides the risk into Low, Medium, High risk of failure.
RECORD_COUNT	Integer	Total number of records checked
DQ_DIMENSION	String	Categorical. Data quality dimensions
DOMAIN	String	Categorical. 6 different Domains (so far)
PROCESS	String	Categorical variable representing the key process the data is used for.
CDE_NAME**	String	Critical Data Element causing the issue*

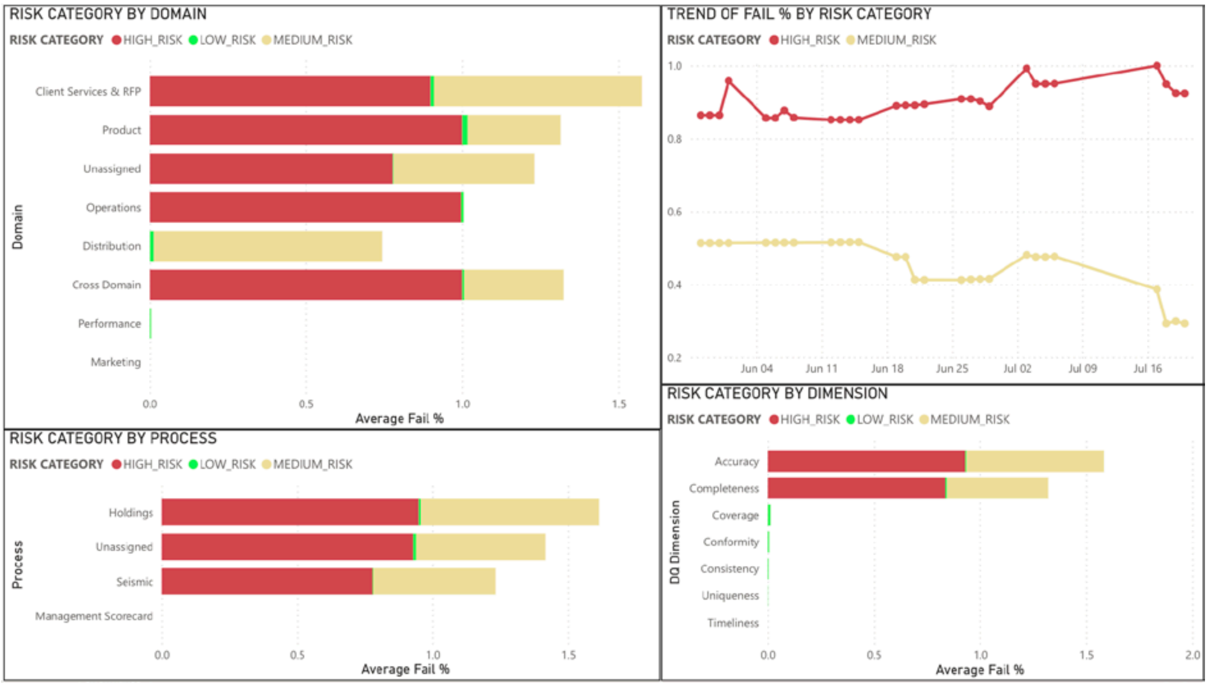
- *RISK CATEGORY – A calculated column that divides the risk of failure into 3 categories: LOW (0 – 25% Average FAIL_PCT), MEDIUM (25 to 75%), HIGH RISK (75% to 100%). The rationale for creating this column is explained under the heading THE BUCKET CHALLENGE.

- ** CDE_NAME was introduced only in the last dataset received (RESULTS 07 07 23 RIT.xlsx), which contained the name of the CDE causing the quality issue. Though we infer CDE_NAME is derived from the ‘CHECK_COLUMN’, we do not have the exact workings of the derivation. Therefore, in our dashboard we have used the ‘CHECK_COLUMN’ column in the place of CDE_Name, for the reason of unavailability of data in the CDE_NAME column, in the earlier datasets. In future, however, CDE_NAME can be used in the place of CHECK_COLUIMN.

THE ART

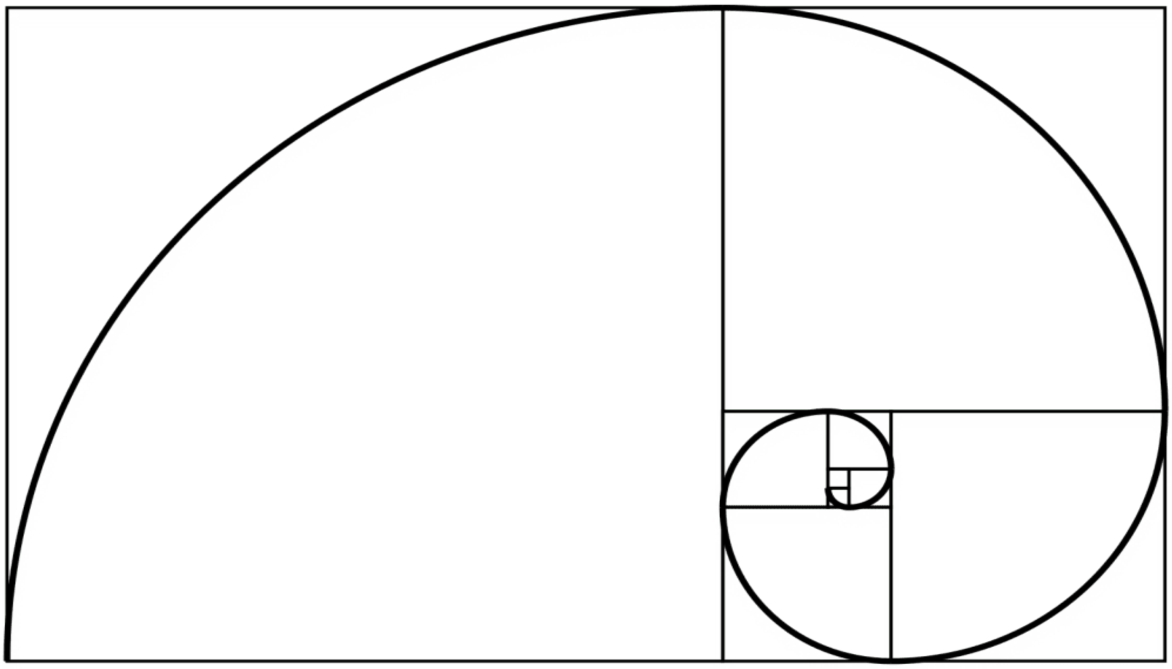
Designing a dashboard is as much an art as a science. Once the objectives were understood clearly, the problem was then broken down into implementable chunks. In particular, we have focussed on the following:

- <u>THE DESIGN ELEMENTS</u>

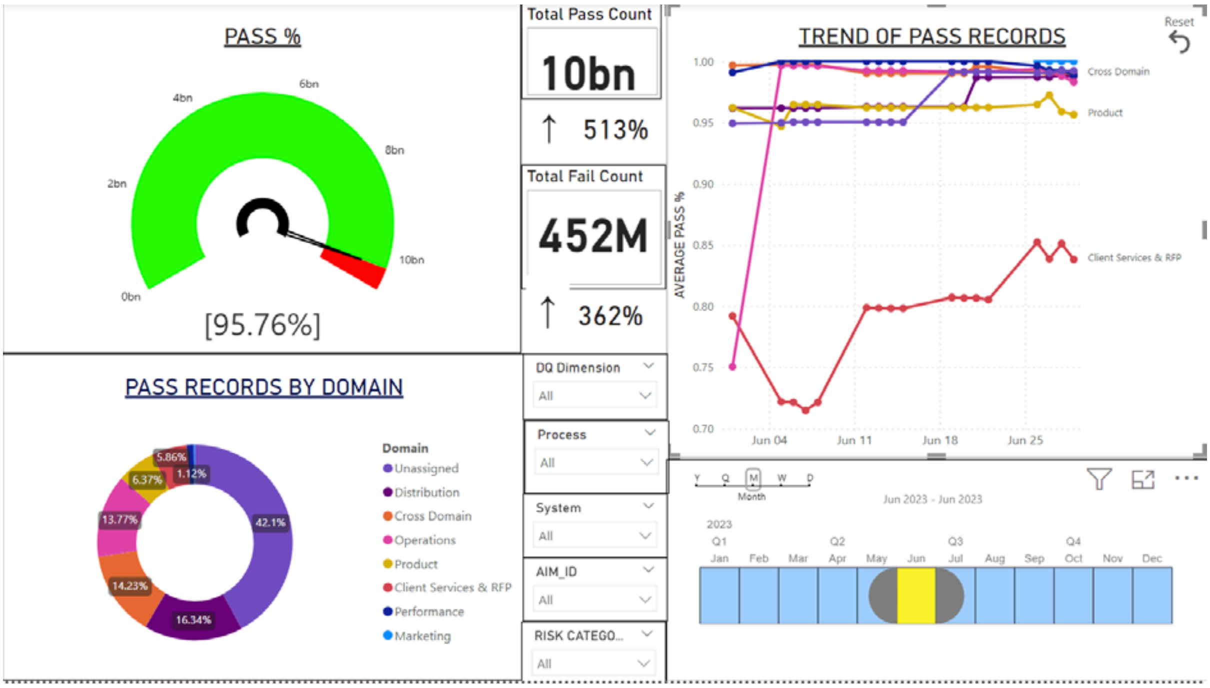


Throughout our dashboard, we have used a consistent colour scheme, that helps the user to focus on the key elements of that page or visual. For example, the ‘blue’ for the domain Client Services & RFP, is used consistently throughout the dashboard for that domain, so that the user can associate that particular colour with that variable. The same principle of consistency is applied to all the design elements like the fonts, font size, charts etc.

- <u>THE GOLDEN RATIO</u>



The golden ratio, also known as the divine proportion, is a special number (equal to about 1.618) that appears many times in geometry, art, and architecture. The golden ratio is found when a line is divided into two parts such that the whole length of the line divided by the long part of the line is also equal to the long part of the line divided by the short part of the line. In other words, it is a guiding technique to divide a rectangular space in proportions that is intuitively appealing to the human eye. Masters like Michelangelo, Da Vinci, architects like Le Corbusier have all used the Golden Ratio to divide their frames. We have attempted to follow the Golden Ratio, to find the right proportions for our visuals within our dashboard.



DQ Dimension
All

Process
All

System
All

AIM_ID
All

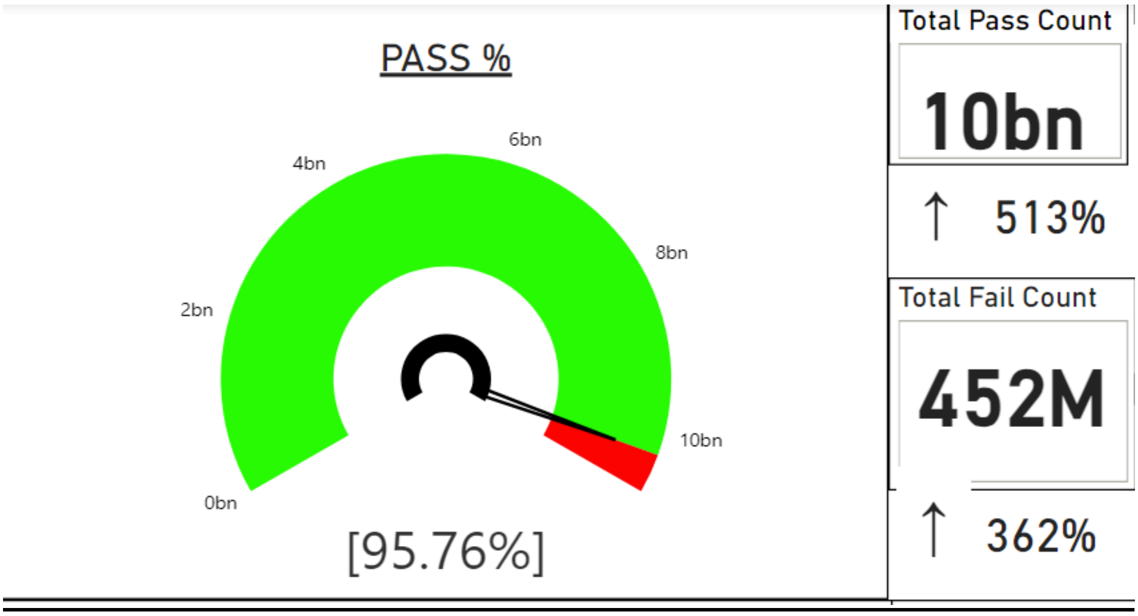
RISK CATEGO...
All

THE SCIENCE

The Science of our dashboard is actually the Tools we have developed to achieve the main objectives:

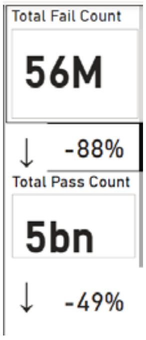
1. A feature/ tool to explain the overall performance of the data across in terms of FAIL & PASS records.
2. A feature/tool to see the Location, Magnitude, and the Trend of the errors.
3. A feature/tool to do a Root-Cause Analysis and derive insights.

TOOL#1: THE TACHOMETER



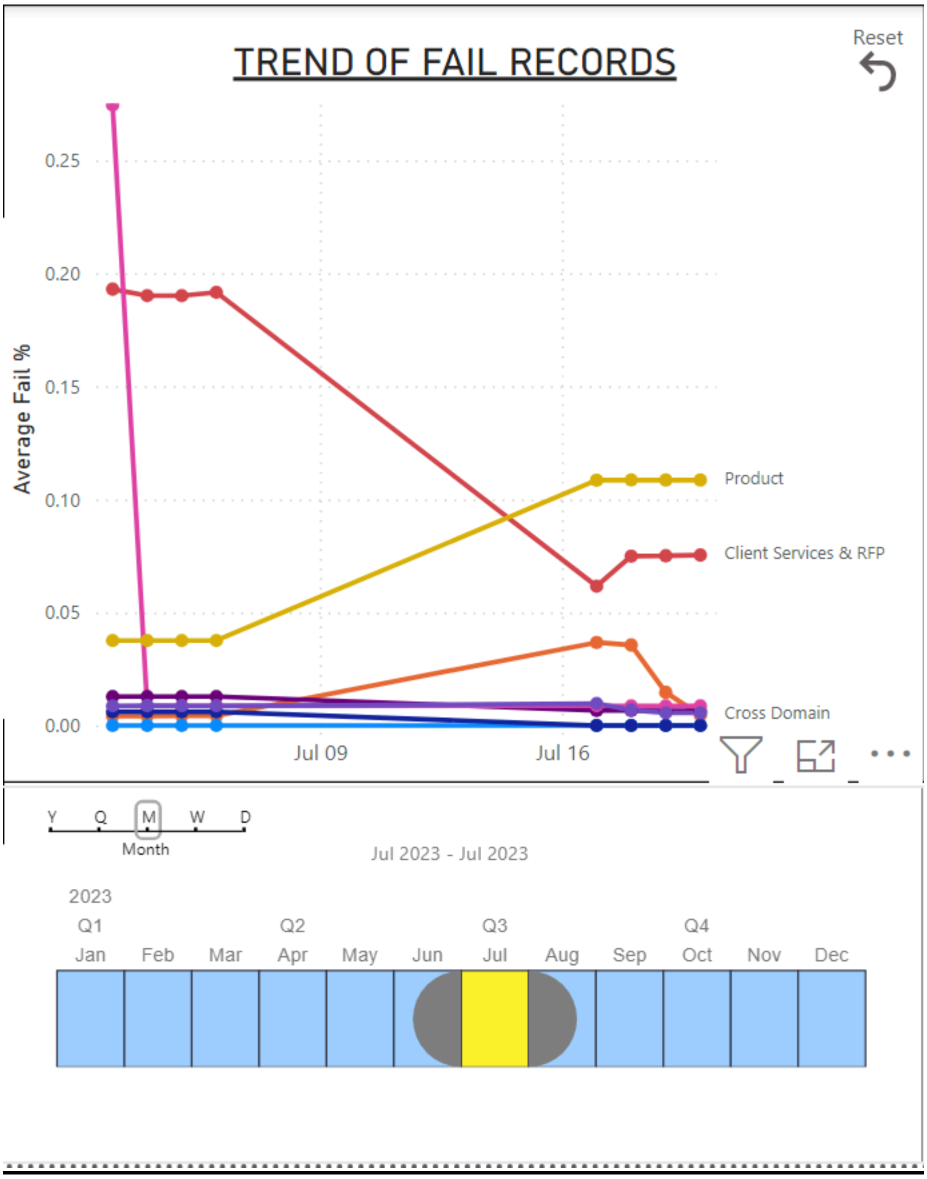
A ‘Gauge’ is the most appropriate visual to explain the overall performance of something at a glance. We have decided to use the Tachometer as a gauge, which gives in an instant, how the data is performing overall, for the selected period. In the above visual, we can see the PASS % as 95.76. This is for the month of July 2023. Similarly, from the FAIL RECORDS page, one can access the FAIL %.

TOOL#2 THE KPI CARDS



The KPI cards accompanying the tachometer, are meant give the overall performance at a glance. It gives you the Total Fail Count in number, and the percentage change, compared with the previous period. In the above example, the cards show the FAIL COUNTS and the PASS COUNTS (the order flips depending whether the user is on the FAIL RECORDS page or PASS RECORDS page), along with the % change. The ‘% change’ card is a great way to compare the performance. Without this feature, it is difficult to gauge if the displayed number is positive or negative, in terms of performance.

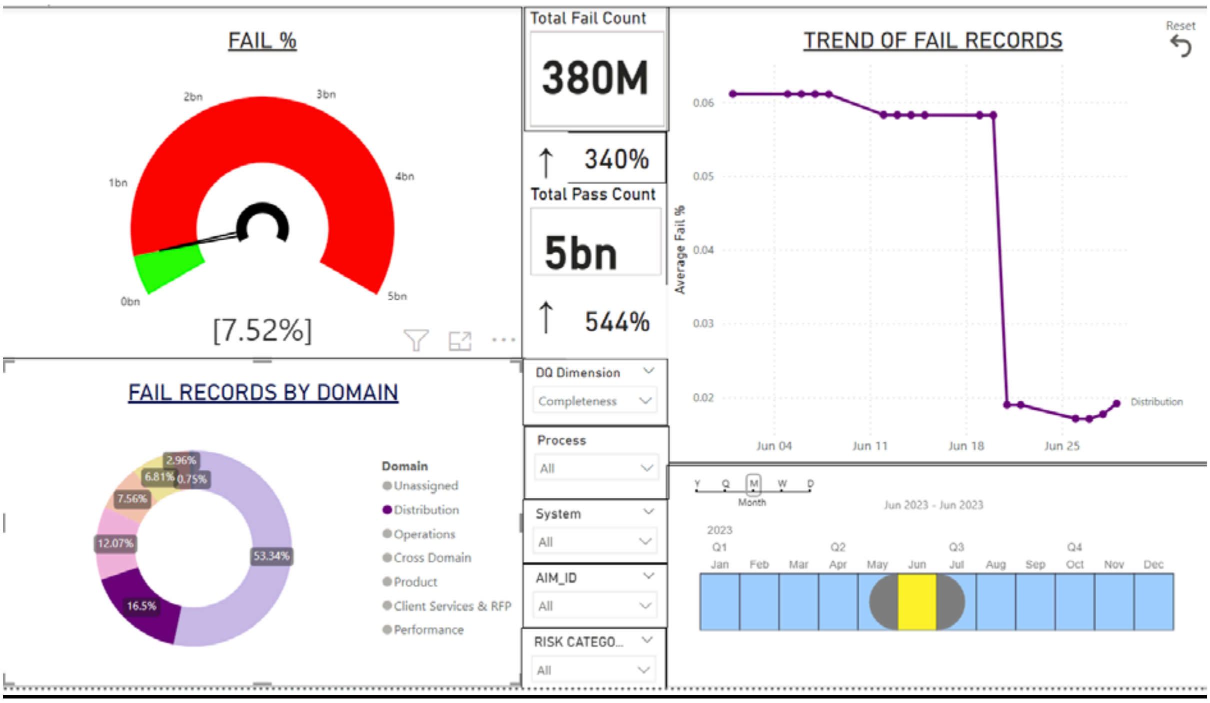
TOOL#3 THE TIMELINE SLICER



Slicers are quintessential part of a dashboard. We could cram a page of a dashboard with ten different charts, but a user can look at only one at a time. Guided by this principle, we have included some smart slicers, that act as windows to the vast expanse of the data. With our carefully placed slicers, a business steward can access the length and breadth of the data, by selecting the buttons in our slicers.

All the data we look at should have time element, and conversely, we should be able to see all the data through a flexible Time point of view. Our dynamic slicer Timeline 2.4.0 achieves this perfectly. A user can select a time period (Year, Quarter, Month, Week, Day) and derive time-based intelligence. It is dynamic, in the sense, it when you change the Time Period, it changes the charts in the entire dashboard automatically. The above visual shows Trend of FAIL RECORDS by DOMAIN.

TOOL#4 THE DONUT



Total Fail Count

380M

↑ 340%

Total Pass Count

5bn

↑ 544%

DQ Dimension

Completeness

Process

All

System

All

AIM_ID

All

RISK CATEGO...

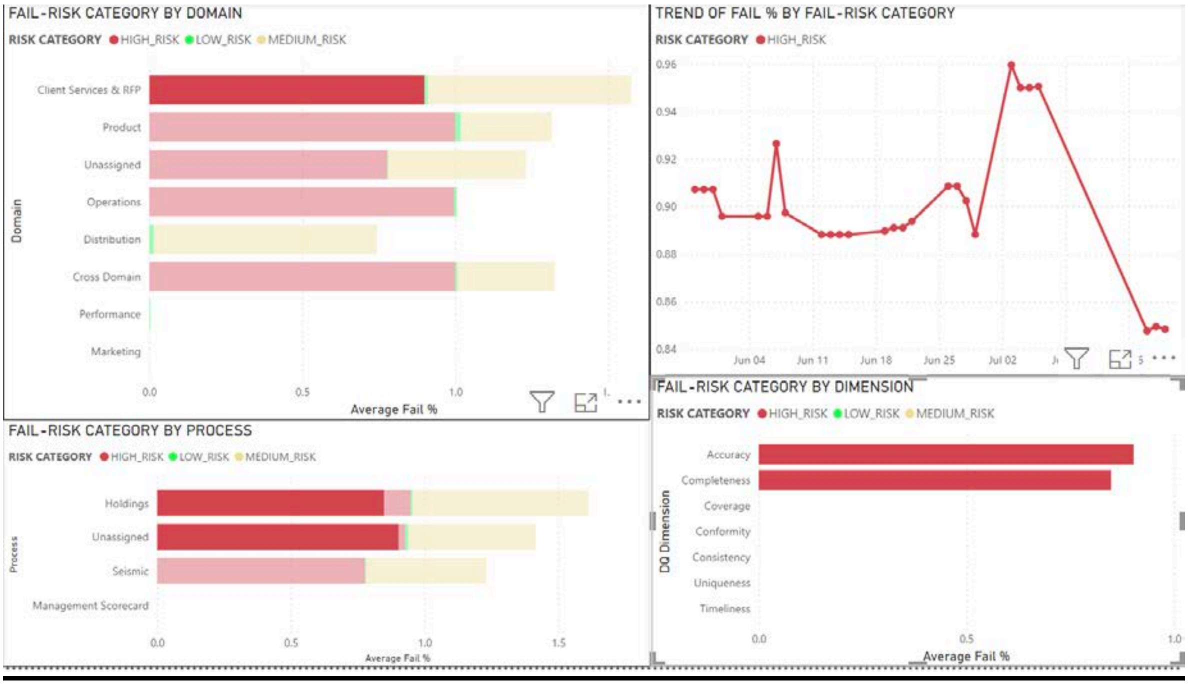
All

Now that we have an idea of the overall performance of the data through the Tachometer and the KPI cards, we can explore in detail the breakdown of the PASS & FAIL RECORDS, through DOMAIN, DIMENSION, PROCESS point of views. The best visual to represent the distribution is a Donut chart. Our DONUT charts give the performance of the data by DOMAIN, by default. One can choose from the slicer on the right of the Donut chart, to select one or options from DIMENSION, PROCESS, SYSTEM, AIM ID or from RISK CATEGORY. In the example above, the Fail Records of the domain ‘Distribution’, of ‘Completeness’ Dimension, for the month of June. We can see the FAIL % (7.52) along with the trend of these ‘Distribution-Completeness’ combination of errors.

TOOL# 5 FAIL-RISK CATEGORY

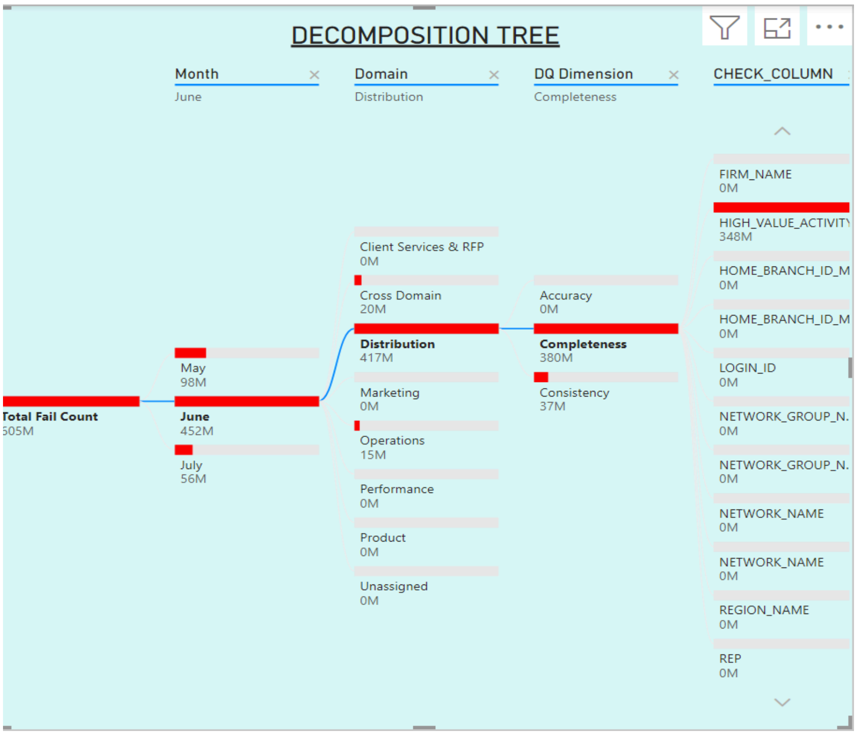
We believe, there are two ways of looking at the data: 1. From a FAIL % (percentage of Failed Records out of Total Records) point of view. 2. From a FAIL COUNT point of view. Choosing one view over the other can result in some erroneous conclusions. For example, if there are two Domains, Domain A and Domain B. And if Domain A has just 100 records, out of which 99 failed, the FAIL % is 99%. Whereas Domain B has, for example, 10 million records and 6 million out of them have failed, giving it a FAIL % of 60. If our tools are designed to rank Domains based on FAIL %, we will end up fixing the 99 failures of Domain A, instead of the 6 million failed records from Domain B. At the same time, we need to investigate why 99% of the records from a particular domain is failing. Therefore, the dashboard should offer the flexibility to analyse the records from BOTH the angles.

Though the main line of investigation should be from the Total Number of Records or ‘Count’ (PASS OR FAIL) point of view, there should be option to investigate from FAIL or PASS % angle. In order to do that, we have designed the ‘FAIL CATEGORY’ page, where we have categorized the FAIL% and PASS% into 3 categories: LOW (0-25%), MEDIUM (25-75%), and HIGH RISK (75% and above) of FAIL %. This will help us identify exactly those Domains,



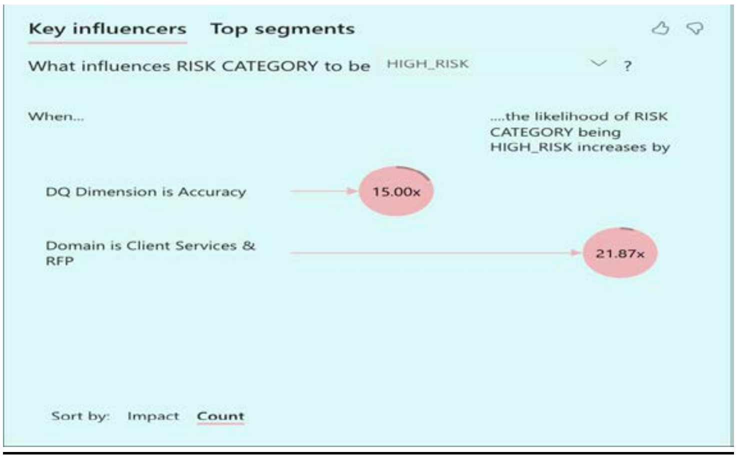
Dimensions, Process etc with HIGH-RISK of failures. Without a bracketing technique, for example, a Domain which consistently produces a 10% average Fail Records, has the possibility of being classed under the same category as another Domain with 99% Average Fail %. With risk category colour-coded (Green, Yellow, Red for LOW, MEDIUM, HIGH RISK, respectively) in our RISK CATEGORY page, a business steward can prioritise their energies towards High-Risk Domains, Dimensions, Datasets etc. In the example below, by clicking on the red slab of Client Services & RFP, one can see the Processes and Dimensions that are causing the high levels of errors, along with its trend.

TOOL# 6 THE DECOMPOSTION TREE



The go-to tool for a Root-Cause Analysis is the Decomposition. It lets you dissect the data in whichever way you want. For example, the above tree traces the Fail_Records in the month of June, in the 'Distribution' domain, for 'Completeness' Dimension, under the Column 'HIGH_VALUE_ACTIVITY'. The bright-red coloured bars conveniently guide the user towards 'problem areas. Within seconds, we can travel down the hierarchy from TOTAL FAIL COUNT to the depth of the SQL commands that are causing the problem. We can do the same for PASS COUNTS, if we choose the option 'Low Value' (of Fail Count).

TOOL# 7 KEY INFLUENCERS



What influences RISK CATEGORY to be HIGH_RISK

When...

....the likelihood of RISK CATEGORY being HIGH_RISK increases by

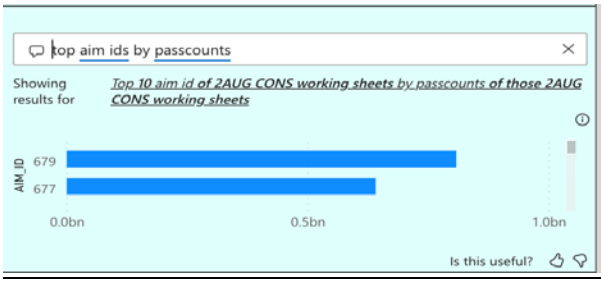
DQ Dimension is Accuracy — 15.00x

Domain is Client Services & RFP — 21.87x

In addition to the Decomposition Tree, the Key Influencers tool analyses the data provided and outputs key crystallized insights. In the example above, the insights produced are as follows:

- The likelihood of RISK CATEGORY being HIGH_RISK (above 75% Fail_PCT) increases 15 times if the Dimension is 'Accuracy'.
- The likelihood of RISK CATEGORY being HIGH_RISK increases 21 times under the Domain 'Client Services & RFP'.

TOOL# 8 Q&A



top aim ids by passcounts

Showing results for *Top 10 aim id of 2AUG CONS working sheets by passcounts of those 2AUG CONS working sheets*

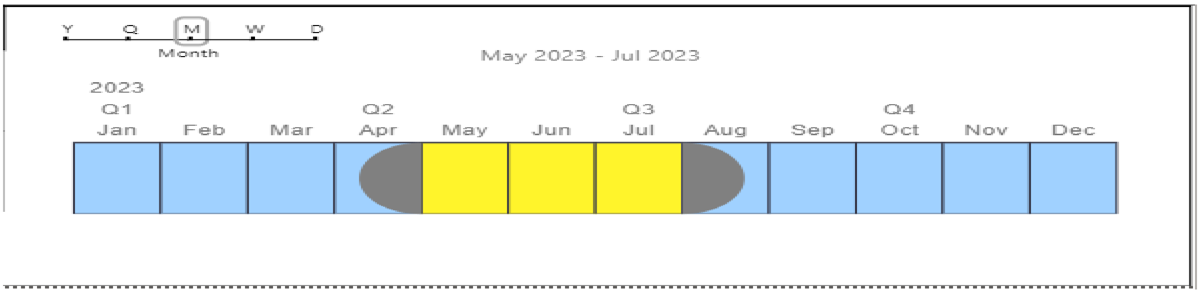
We have also included a Q&A box, that produces instant charts and tables depending on the complexity of the question. Its sophisticated question box with predictive texting and auto-spell Check, makes it a useful tool to produce insightful graphs and charts.

DASHBOARD DEMO

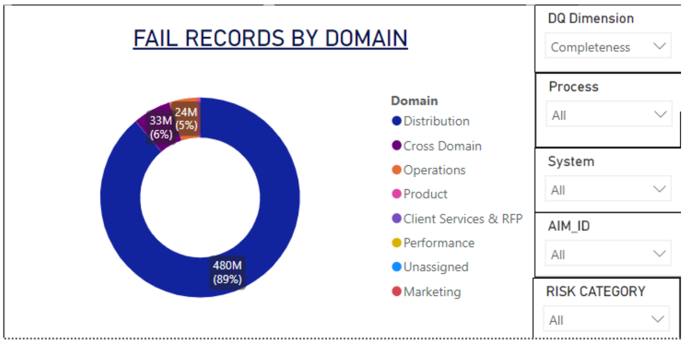
In this section we are going to demonstrate step-by-step, how to locate millions of failed records within seconds. By following the steps below root-cause of the failures can be located in seconds, rather than minutes. For example;

TIMESLICER>FAILDATA>DECOMPOSITION TREE> DOMAIN>SQL COMMAND

1. **SELECT THE TIME PERIOD**: In this example we choose months May, June, and July 2023 (the period for which the data is available).



2. From the **FAIL DATA page**, see which DOMAIN has the greatest number of FAIL RECORDS or biggest share of the donut. In this case, Distribution (480M, 89% of failures)



DQ Dimension
Completeness

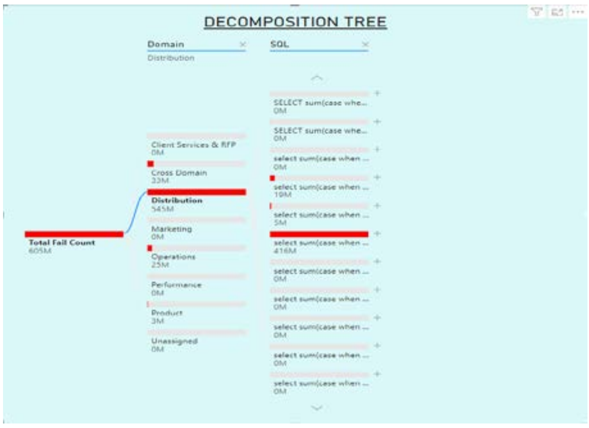
Process
All

System
All

AIM_ID
All

RISK CATEGORY
All

3. INSIGHTS PAGE>DECOMPOSTION TREE> DOMAIN>DISTRIBUTION>SQL



TOPLINE INSIGHTS

- DOMAIN: 90% (545M out of 605M) of failures occur under ‘Distribution’.
- DIMENSION: 89% (540M out of 605M) happen to be ‘Completeness’ mismatches.
- PROCESS: 95% (572M out of 605M) does not have any Process assigned to it (Unassigned).
- MONTH: Most number of errors have occurred during June 2023, with 41M Fail Records on 19th June.
- DATASETS: 3 Datasets are responsible for bulk of the errors (PUBLIC.VW_FACT_ACTIVITY (405M), PCARW. VW_FACT_ACTIVITY (104M) PUBLIC.VW_FACT_HOLDING_EQ_MONTHLY (55M)
- SYSTEM: ALL the errors are from ARW
- COLUMN: HIGH_VALUE_ACTIVITY in the dataset PUBLIC.VW_FACT_ACTIVITY is responsible for 72 % (440M out of 605M) of errors.
- SQL: This particular SQL command is behind 69% (416M out of 605M) errors:

```
select sum(case when HIGH_VALUE_ACTIVITY is NULL then 0 else 1 end)/count(1) as Pass_PCT, sum(case when HIGH_VALUE_ACTIVITY is NULL then 1 else 0 end)/count(1) as Fail_PCT, sum(case when HIGH_VALUE_ACTIVITY is NULL then 0 else 1 end) as Pass_Count, sum(case when HIGH_VALUE_ACTIVITY is NULL then 1 else 0 end) as Fail_Count, count(1) as Records_Tested from "PCARW"."PUBLIC"."VW_FACT_ACTIVITY";
```

In summary, this SQL command generates various metrics related to the presence of non-NULL and NULL values in the column HIGH_VALUE_ACTIVITY within the "VW_FACT_ACTIVITY" table. It provides pass and fail percentages along with corresponding counts for both scenarios and also the total number of records tested. In almost all the cases (605M of them), the errors are happening in columns, there are NULL values, where they are not supposed to be. A possible solution to the problem could be, ensuring the condition “cannot be null”, is met, at the time of Table creation.

CONCLUSION

We started off by defining our objectives:

1. A feature/ tool to explain the overall performance of the data across in terms of FAIL & PASS records.
2. A feature/tool to see the Location, Magnitude, and the Trend of the errors.
3. A feature/tool to do a Root-Cause Analysis and derive insights.

To achieve those objectives, we have created the afore-mentioned tools and have deployed them in an aesthetically pleasing way. The tools we developed will give the user a quick understanding of the data and answers to some persistent questions. And finally using our tools, we have demonstrated how a user can locate and fix the cause of failures.