Problem:

Two systems with large amounts of customer data may not communicate properly 100% of the time and therefore will need their data reconciled. It’s important to keep our data clean between the systems because we don’t want to have the most accurate numbers when analyzing data as well as clean data for customer contact.

Proposed Solution:

Create a R Script that allows us to flag down the possible mismatched combinations between the two systems and target those customer columns for clean up. By automating this process, analysts no longer have to spend hours combing over thousands of lines of data but instead can run this efficient and effective script to produce a better outcome. This will allow us to run a reconciliation on a more frequent basis as well.

Findings:

Within the data I conducted some analysis that would be useful to report and highlight for the clean up in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Flag | Count | % | Validation | Notes |
| Enrolled in dataset1, Unenrolled in dataset2 | 1 | 2% | Yes | Source of truth being dataset2 (SeeLoad system)  To unenroll from dataset1(Franklin system) |
| Enrolled in dataset2, Unenrolled in dataset1 | 3 | 6% | Yes | Source of truth being dataset2 (SeeLoad system)  To enroll from dataset1(Franklin system) |
| Install Date mismatch | 1 | 2% | Yes | Source of truth being dataset2 (SeeLoad system)  To change date from dataset1(Franklin system) |
| Mismatching Names | - | - | N/A |  |
| Mismatching Address | - | - | N/A |  |
| AccountID mismatches | 3 | 6% | Yes | We see that there are 3 AccountIDs in Franklin that aren’t in SeeLoad |
| SAID mismatches | 6 | 12% | Yes | We see that there are 3 SAIDs in Franklin that aren’t in SeeLoad and 3 SAIDs in SeeLoad that aren’t in Franklin |
| DRUID mismatches | 3 | 6% | Yes | We see that there are 3 DRUIDs in SeeLoad that aren’t in Franklin |
| SPID mismatches | 3 | 6% | Yes | We see that there are 3 SPIDs in SeeLoad that aren’t in Franklin and 4 SPIDs in Franklin that aren’t in SeeLoad |
| Serial Number mismatch | 1 | 2% | Yes | We see that there is 1 Serial Number that needs to be matched to SeeLoad |
| Complete match | 24 | 48% | Yes | Within the two datasets, we’ll be looking at the other 52% of the data |

Conclusion:

Using the flag/bucket structure, we can pinpoint exactly what needs to be cleaned up in which system/dataset. Knowing the source of truth being dataset2(SeeLoad system), we can then change the data accordingly in batches using the files from the results output folder.