Data Migration Quality Analysis

Presented by Jeffrey Benson

Agenda

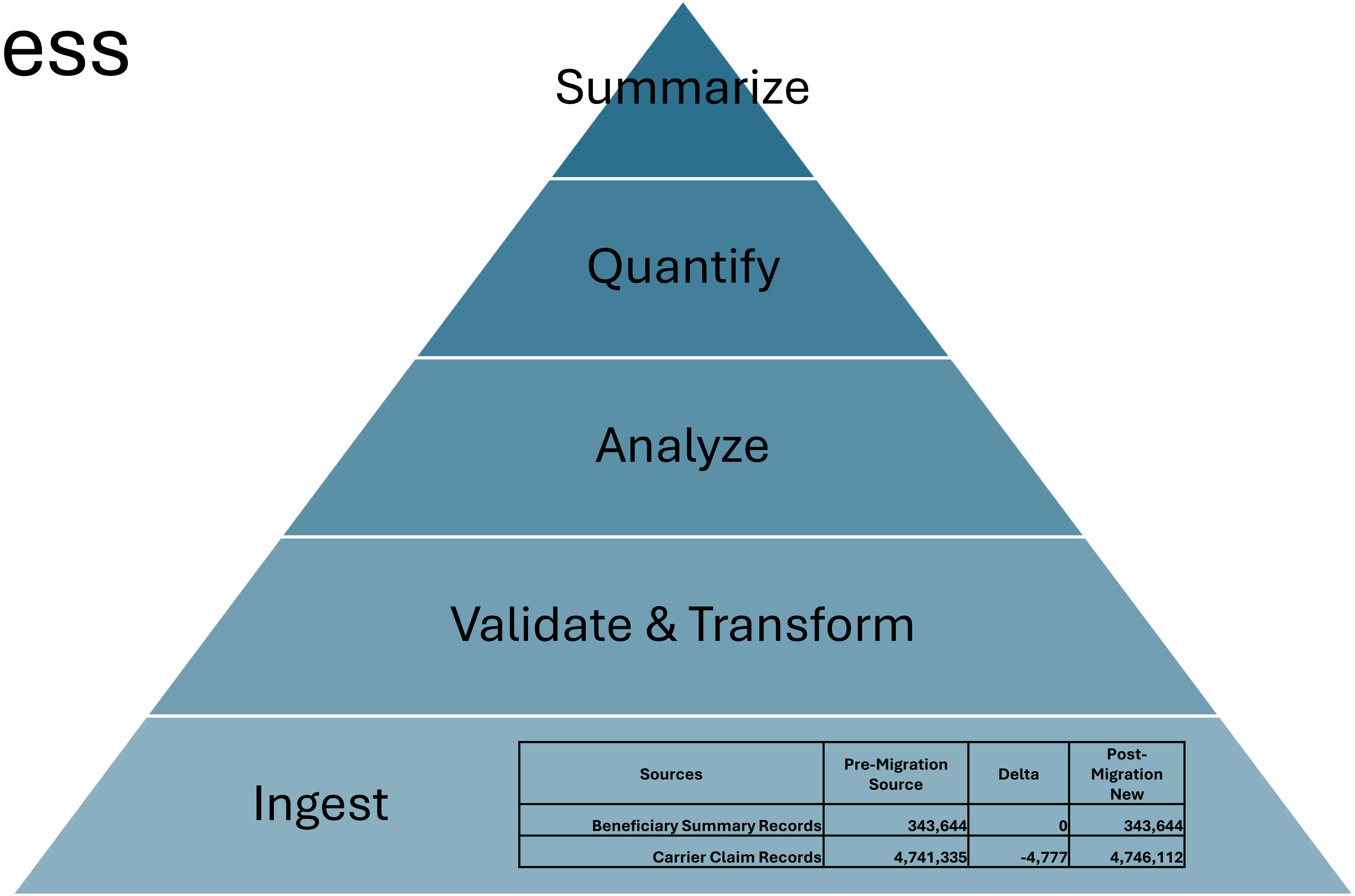
01 Process

02 Findings

03 Next Steps

04 Tech Stack

Process



Sources	Pre-Migration Source	Delta	Post-Migration New
Beneficiary Summary Records	343,644	0	343,644
Carrier Claim Records	4,741,335	-4,777	4,746,112

Capability of Migration Process

Data Set	Total Rows	Defective Data Elements	Total Data Elements	DPMO*	Process Yield	Sigma
Carrier Claims	474.7K	530.4K	484,150.9K	1095.5	99.89045%	~4.58
Beneficiary Summary	343.8K	8.2K	10,657.9K	768.7	99.92313%	~4.68

“Six Sigma” represents a process capability with a 99.99966% yield, or 3.4 defects per million opportunities.

__Data Migration Process Capability:__

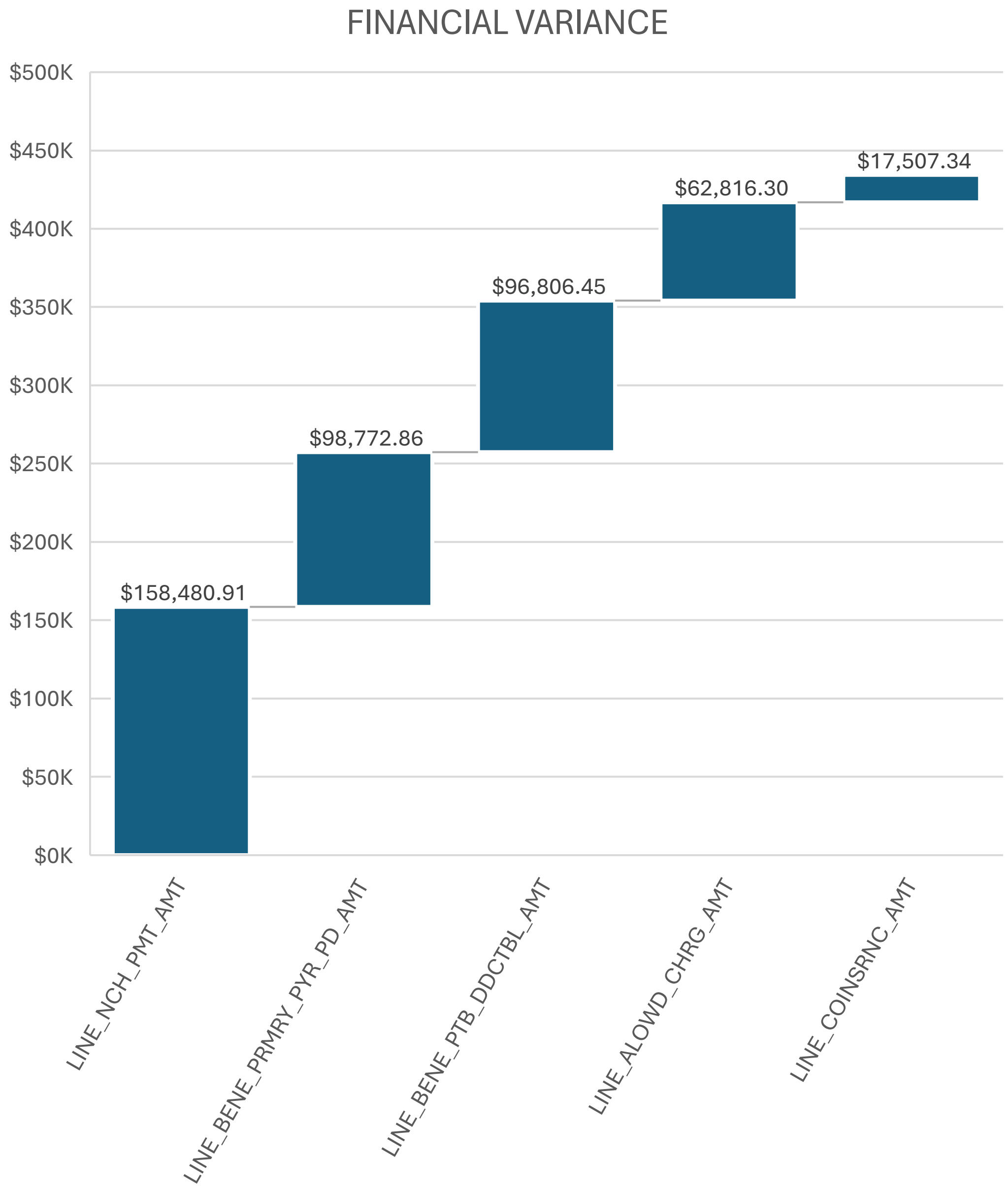
- Carrier Claims: 99.89045% yield, or 1,095.5 defects per million opportunities*.
- Beneficiary Summary: 99.92313% yield, or 768.7 defects per million opportunities*.

Pareto Analysis of Claims Data Defects

SOURCE	DATA SET	TOTAL DEFECTS	RUNNING PCT	DPMO	SIGMA LEVEL
Carrier Claims	LINE_PRCSG_IND_CD	75,497	13.93%	15,906	3.656
Carrier Claims	LINE_ICD9_DGNS_CD	75,238	27.82%	15,851	3.657
Carrier Claims	TAX_NUM	75,176	41.70%	15,838	3.658
Carrier Claims	PRF_PHYSN_NPI	75,169	55.57%	15,836	3.658
Carrier Claims	HCPCS_CD	75,106	69.43%	15,823	3.658
Carrier Claims	ICD9_DGNS_CD	46,708	78.05%	9,840	3.844
Carrier Claims	LINE_NCH_PMT_AMT	31,709	83.91%	6,680	3.988
Carrier Claims	LINE_BENE_PTB_DDCTBL_AMT	20,828	87.75%	4,388	4.137
Carrier Claims	LINE_BENE_PRMRY_PYR_PD_AMT	20,639	91.56%	4,348	4.141
Carrier Claims	LINE_ALOWD_CHRG_AMT	17,531	94.80%	3,693	4.197
Carrier Claims	LINE_COINSRNC_AMT	16,775	97.89%	3,534	4.211
Carrier Claims	CLM_FROM_DT	5,707	98.95%	1,202	4.553
Carrier Claims	CLM_THRU_DT	5,707	100.00%	1,202	4.553

Financial Impact of Claims Data Defects

DATA SET	FINANCIAL VARIANCE	RUNNING PCT OF TOTAL
LINE_NCH_PMT_AMT	$ 158,480.91	36.48%
LINE_BENE_PRMRY_PYR_PD_AMT	$ 98,772.86	59.22%
LINE_BENE_PTB_DDCTBL_AMT	$ 96,806.45	81.51%
LINE_ALOWD_CHRG_AMT	$ 62,816.30	95.97%
LINE_COINSRNC_AMT	$ 17,507.34	100.00%
TOTAL	**$ 434,383.86**	**100.00%**



Pareto Analysis of Beneficiary Data Defects

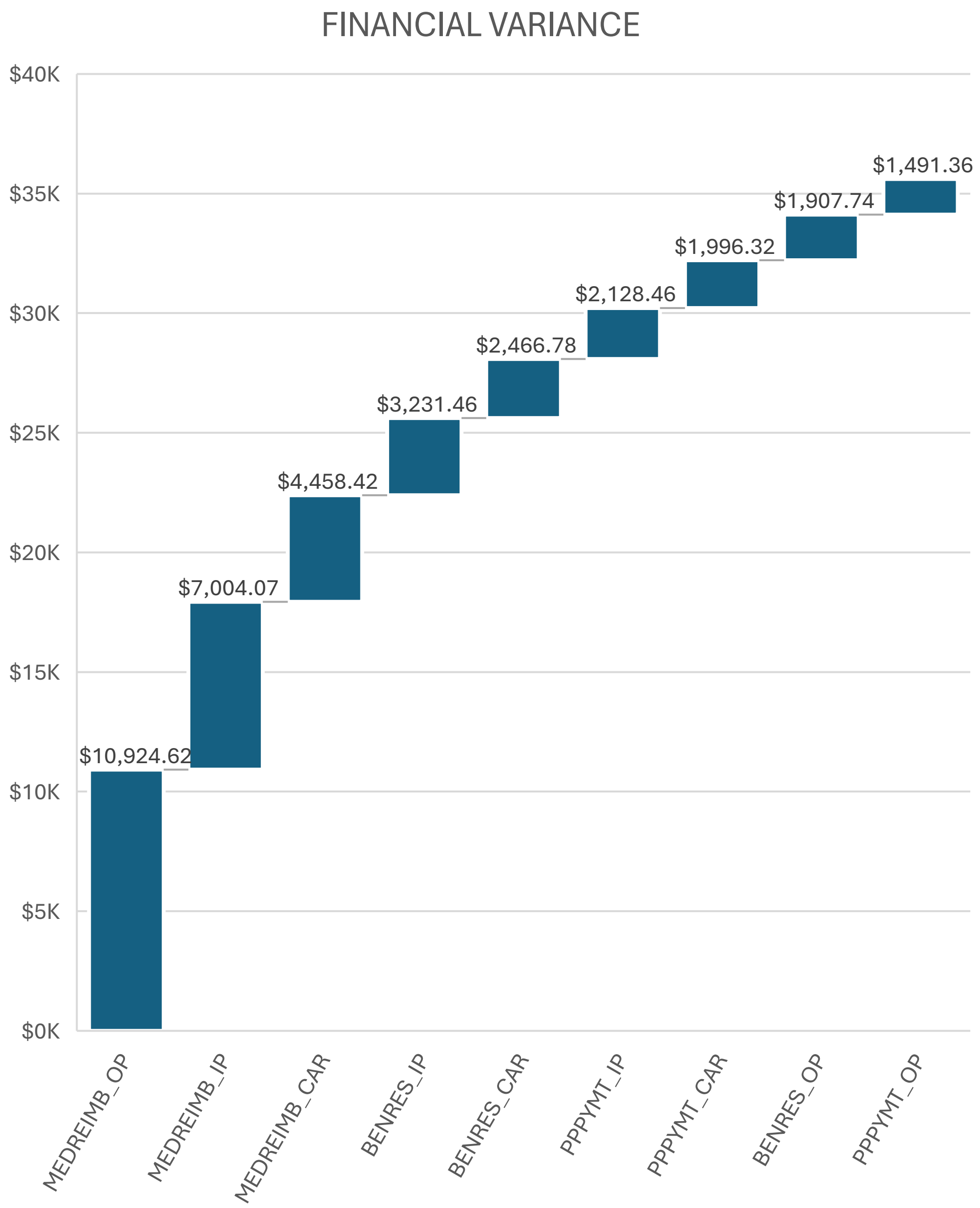
SOURCE	DATA SET	TOTAL DEFECTS	RUNNING PCT	DPMO	SIGMA LEVEL
Beneficiary Summary	BENE_BIRTH_DT	496	6.05%	1,443	4.498
Beneficiary Summary	BENE_HI_CVRAGE_TOT_MONS	337	10.17%	980	4.613
Beneficiary Summary	BENE_SMI_CVRAGE_TOT_MONS	322	14.10%	937	4.626
Beneficiary Summary	BENE_COUNTY_CD	318	17.98%	925	4.629
Beneficiary Summary	BENE_DEATH_DT	318	21.86%	925	4.629
Beneficiary Summary	BENE_ESRD_IND	318	25.74%	925	4.629
Beneficiary Summary	BENE_RACE_CD	318	29.62%	925	4.629
Beneficiary Summary	BENE_SEX_IDENT_CD	318	33.50%	925	4.629
Beneficiary Summary	SP_ALZHDMTA	318	37.39%	925	4.629
Beneficiary Summary	SP_CHF	318	41.27%	925	4.629
Beneficiary Summary	SP_CHRNKIDN	318	45.15%	925	4.629
Beneficiary Summary	SP_CNCR	318	49.03%	925	4.629
Beneficiary Summary	SP_COPD	318	52.91%	925	4.629
Beneficiary Summary	SP_DEPRESSN	318	56.79%	925	4.629
Beneficiary Summary	SP_DIABETES	318	60.67%	925	4.629
Beneficiary Summary	SP_ISCHMCHT	318	64.56%	925	4.629
Beneficiary Summary	SP_OSTEOPRS	318	68.44%	925	4.629
Beneficiary Summary	SP_RA_OA	318	72.32%	925	4.629
Beneficiary Summary	SP_STATE_CODE	318	76.20%	925	4.629
Beneficiary Summary	SP_STRKETIA	318	80.08%	925	4.629
Beneficiary Summary	PLAN_CVRG_MOS_NUM	289	83.61%	841	4.657
Beneficiary Summary	BENRES_CAR	272	86.93%	791	4.674
Beneficiary Summary	MEDREIMB_CAR	260	90.10%	756	4.687
Beneficiary Summary	MEDREIMB_OP	215	92.73%	625	4.740
Beneficiary Summary	BENRES_OP	195	95.11%	567	4.766
Beneficiary Summary	BENE_HMO_CVRAGE_TOT_MONS	124	96.62%	361	4.886
Beneficiary Summary	PPPYMT_CAR	78	97.57%	227	5.002
Beneficiary Summary	MEDREIMB_IP	68	98.40%	198	5.035
Beneficiary Summary	BENRES_IP	62	99.16%	180	5.057
Beneficiary Summary	PPPYMT_IP	36	99.60%	105	5.182
Beneficiary Summary	PPPYMT_OP	33	100.00%	96	5.201

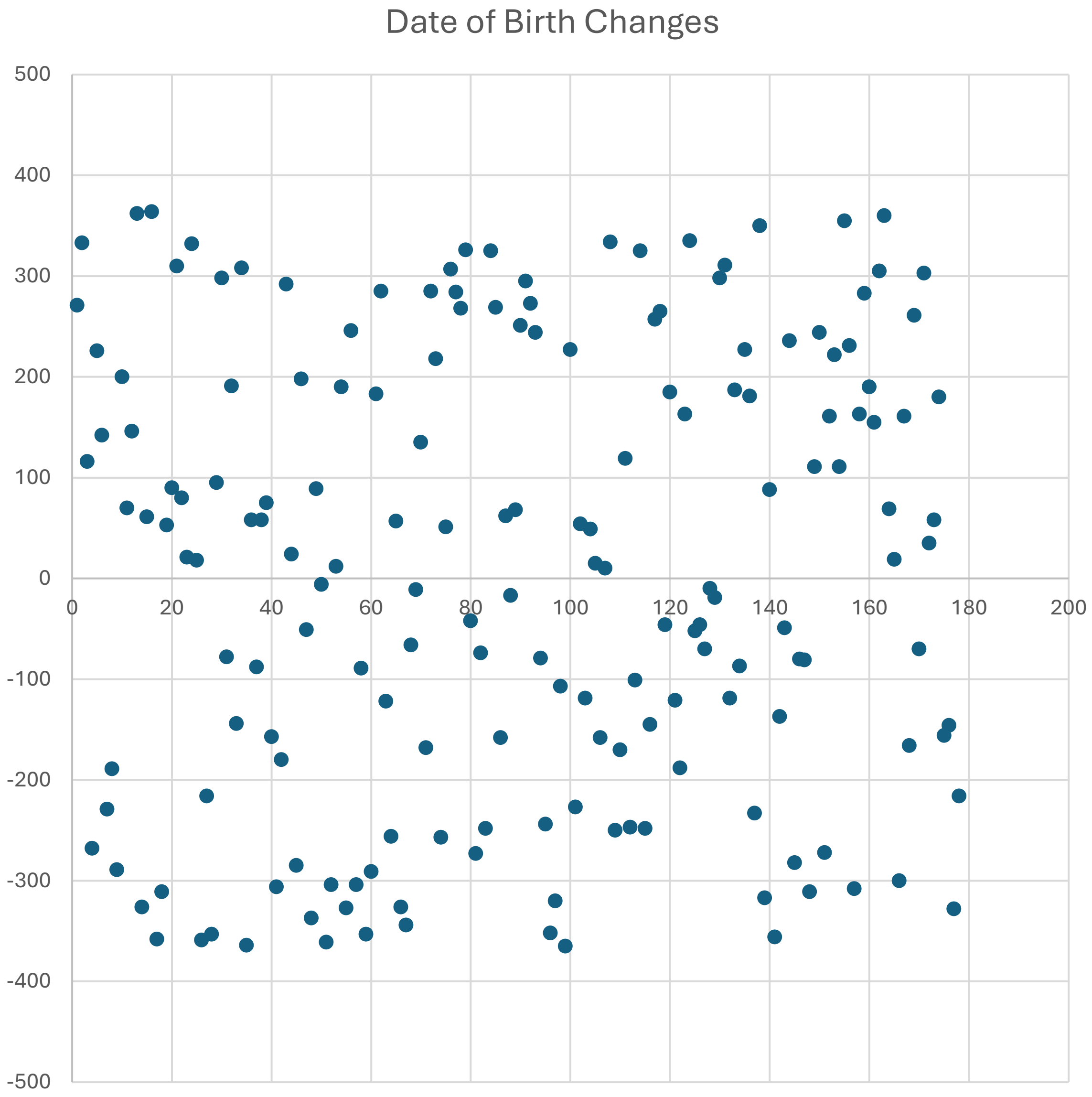
1,322 Beneficiary Records do not match between the Pre- & Post-Migration tables.

One or more data fields are out of sync.

Financial Impact of Beneficiary Data Defects

DATA SET	FINANCIAL VARIANCE	RUNNING PCT OF TOTAL
MEDREIMB_OP	$ 10,924.62	30.68%
MEDREIMB_IP	$ 7,004.07	50.35%
MEDREIMB_CAR	$ 4,458.42	62.87%
BENRES_IP	$ 3,231.46	71.94%
BENRES_CAR	$ 2,466.78	78.87%
PPPYMT_IP	$ 2,128.46	84.85%
PPPYMT_CAR	$ 1,996.32	90.45%
BENRES_OP	$ 1,907.74	95.81%
PPPYMT_OP	$ 1,491.36	100.00%
TOTAL	**$ 35,609.23**	**100.00%**

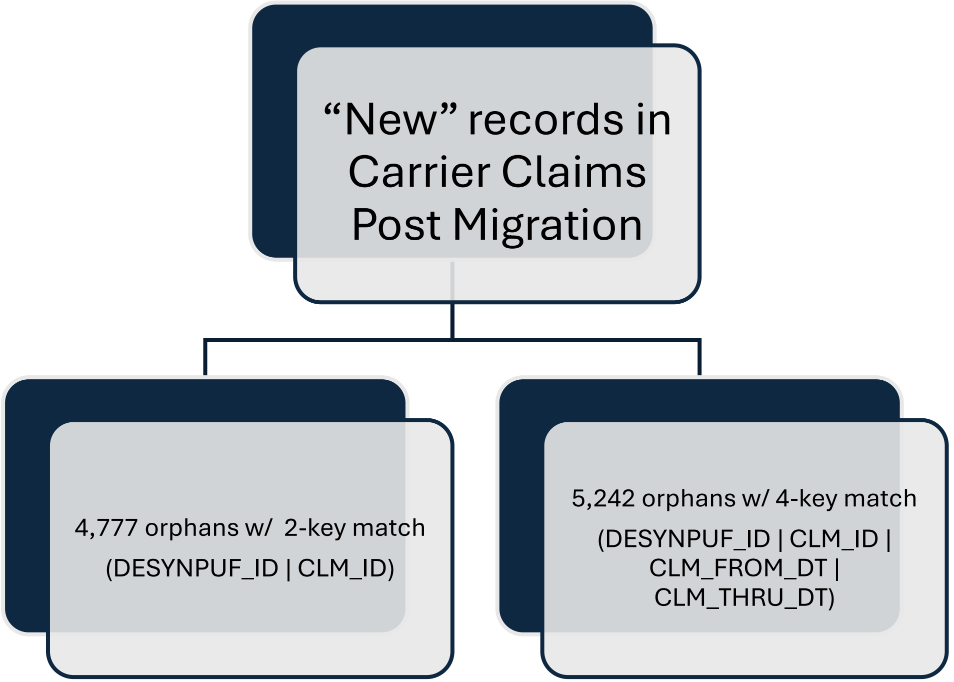




Beneficiary “Date of Birth” Errors

178 Beneficiary “Dates of Birth” changed after the Migration

Carrier Claims Orphan Analysis



“New” records in Carrier Claims Post Migration

4,777 orphans w/ 2-key match

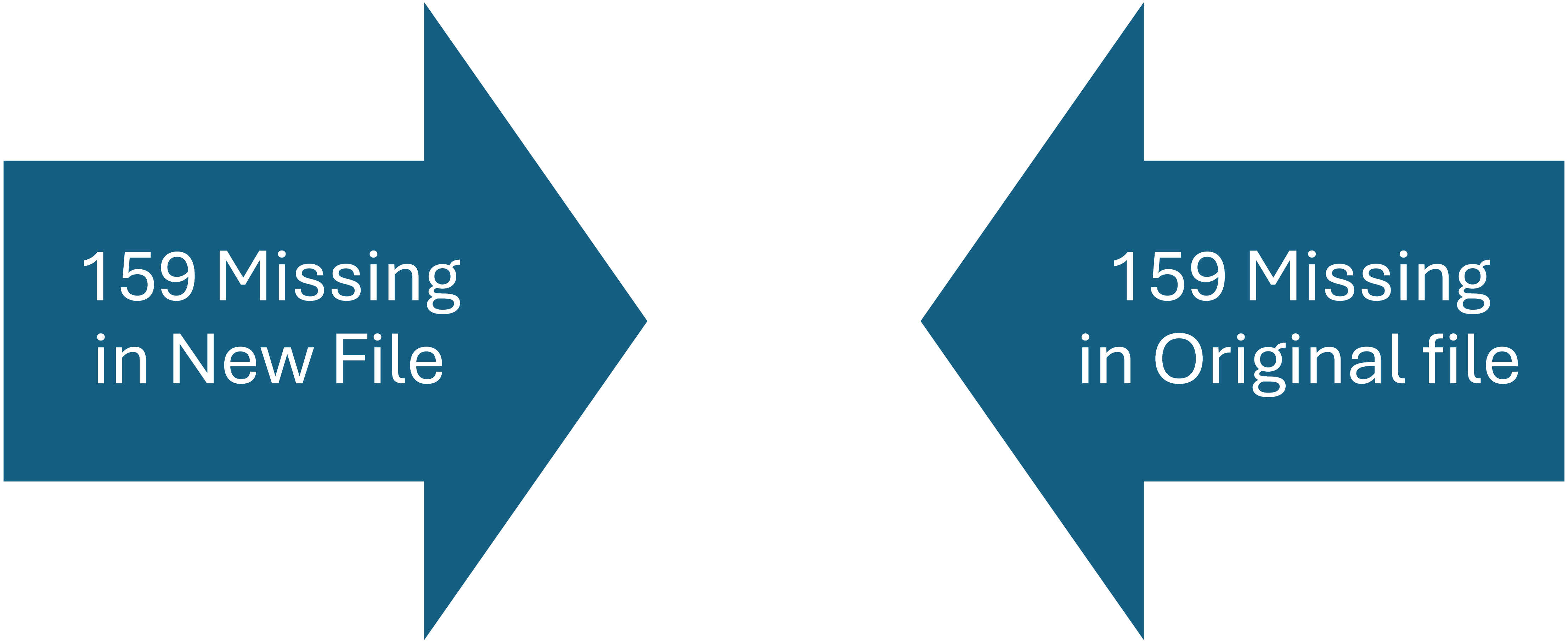
(DESYNPUF_ID | CLM_ID)

5,242 orphans w/ 4-key match

(DESYNPUF_ID | CLM_ID | CLM_FROM_DT | CLM_THRU_DT)

Orphans Increased w/ 4-key match due to defects in CLM_FROM_DT and CLM_THRU_DT Post-Migration

Beneficiary ID Errors



159 Missing in New File

159 Missing in Original file

Need to Investigate: If same persons, why did the IDs change post migration?

Next Steps

Error Datasets Available

Datasets showing every single field that changed after the migration were created and are available to serve the Application and Database Engineers. May help with root case analysis (RCA).

Detailed Quality Scores Available

Detailed Quality Scores available for every field in both Beneficiary and Carrier Claims tables. Helps with setting priorities.

Analysis Positioned to Support Deep Dives

Can provide analysis to answer specific questions about the data migration that this presentation did not answer.

Analytics Improves with Time and Familiarity

Next: Interview stakeholders, gather more context about the application, data, and expectations for the migration. Then update analysis to accommodate new insights and provide custom views to answer questions that stakeholders raise.

Tech Stack Utilized

Languages	AI
Python 3.13 (pandas, sqlalchemy, psycopg2-binary, pytest, tabulate, scipy, numpy, duckdb, duckdb-engine)	Gemini Pro
DuckDB (Database)	Notebook LM
SQL	Perplexity

- Jeffrey designed the queries and set the approach used for the entire pipeline, analytical engine, outputs, etc.
- AI assistance: “Human in the Loop” code syntax, environment set-up , data ingestion, and PDF Summarization.
- Level of Effort / Behind the Scenes:
  - 28 hours invested (4 hours set-up + 24 hours data exploration, analysis, and summarization).
    - 17 py scripts (~1,500 lines of code)
    - 1 sql script (~2,000 lines) to transform / analyze and prepare interim / final views of the data.
- Utilized “report.md” and “report.html” for automated feedback to communicate the data quality.
- Provided PowerPoint presentation to demonstrate how we would present the findings to others in a presentation format.