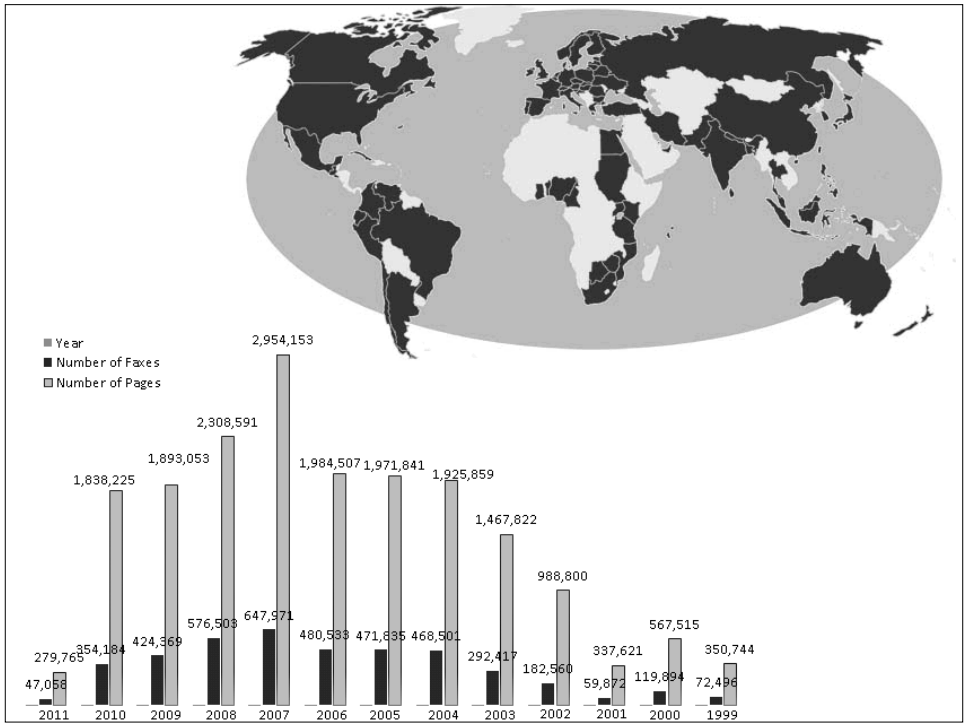
# DataFax Upgrades: Verifying Data Completeness and Correctness

Nidhi Jethoo
Population Health Research Institute
http://www.phri.ca


## The Goals

- Ensure data integrity is maintained through migrations
- Have formal documentation for internal and external auditing
- Have a reproducible automated process




## Why do we do this?

- To ensure study results are not impacted
  - IT processes should never have an impact on results
  - Data completeness and correctness are top priority
- To meet regulatory requirements for:
  - Documentation of process
  - Validation of results
- To ensure study teams have a high degree of confidence in post migration study data

## Definitions

- Consistency
  - Scripted process to ensure common reproducible actions are done to avoid human error
- Completeness
  - All the data exists and is unaltered
- Correctness of the study data
  - Reports on the data say the same before and after migration/upgrade

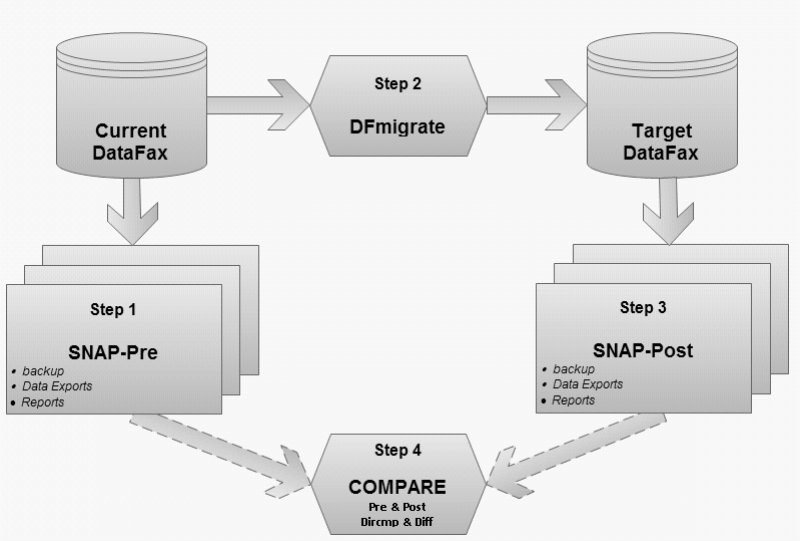
## The Challenge

- How do we know all data was migrated properly and prove it?
- How do we know that the new system is processing the data records correctly?
- How do we ensure that human errors are eliminated or at least kept to the lowest risk possible?

## The Approach

- Script steps to remove human errors
- Log all the steps for review and auditing
- Capture the state of the system before and after the migration/upgrade
- Compare the state of the system after the migration with pre-migration state

## The Migration Process



## The Process

- Do the following pre and post migration
- for $study_num in $studies ; do
  - mkdir SNAP383/$study_num
  - cd SNAP383/$study_num
  - Enable study using DFenable.rpc
  - run DFstudyPerms and save output
  - Run DFexport.rpc for each plate and save output
  - Run DFsas for each plate and save output
  - Compile edit checks using DFcompiler and save output
  - Run specified reports and save output
  - Disable study using DFdisable.rpc
- Done
- Keep the 2 log folders identical to make the comprision easy

## List of Reports

- Schema reports
  - DF_SScenters
  - DF_Ssschema
  - DF_Ssvars
  - DF_Ssvisitmap
- Integrity Checks reports
  - DF_ICkeys
  - DF_ICrecords
  - DF_ICschema
  - DF_ICvisitdates
  - DF_ICvisitmap
  - DF_ICqcs
  - DF_ICcenters
  - DF_ICimages

## Structure of Pre-Migration SNAP logs

```
-bash-3.00$ pwd
/ous2/migration/logs/SNAP383
-bash-3.00$ ls
1    109  119  129  139  152  178  222  232  242  27   37   47   57   67   77   87   97
10   11   12   13   14   153  18   223  233  243  28   38   48   58   68   78   88   98
100  110  120  130  140  155  19   224  234  244  29   39   49   59   69   79   89   99
101  111  121  131  144  16   2    225  235  246  3    4    5    6    7    8    9
102  112  122  132  145  160  20   226  236  249  30   40   50   60   70   80   90
103  113  123  133  146  17   200  227  237  25   31   41   51   61   71   81   91
104  114  124  134  148  170  204  228  238  250  32   42   52   62   72   82   92
105  115  125  135  149  171  205  229  239  251  33   43   53   63   73   83   93
106  116  126  136  15   175  206  23   24   252  34   44   54   64   74   84   94
107  117  127  137  150  176  21   230  240  253  35   45   55   65   75   85   95
108  118  128  138  151  177  22   231  241  26   36   46   56   66   76   86   96
-bash-3.00$ cd 11
-bash-3.00$ ls
11               11.SSschema      11.d11           11.plt002.exp    11.plt016.exp
11.DFedits       11.SSvars        11.d12           11.plt003.exp    11.plt017.exp
11.DFstudyPerms  11.SSvisitmap    11.d13           11.plt004.exp    11.plt018.exp
11.ICcenters     11.d02           11.d14           11.plt005.exp    11.plt019.exp
11.ICimages      11.d03           11.d15           11.plt006.exp    11.plt020.exp
11.ICkeys        11.d04           11.d16           11.plt007.exp    11.plt021.exp
11.ICqcs         11.d05           11.d17           11.plt008.exp    11.plt086.exp
11.ICrecords     11.d06           11.d18           11.plt009.exp    11.plt501.exp
11.ICschema      11.d07           11.d19           11.plt010.exp    11.sas
11.ICvisitdates  11.d08           11.d20           11.plt011.exp
11.ICvisitmap    11.d09           11.d23           11.plt012.exp
11.SScenters     11.d10           11.plt001.exp    11.plt013.exp
-bash-3.00$
```

## Structure of Post-Migration SNAP logs

```
-bash-3.00$ cd ../../SNAP401
-bash-3.00$ ls
1    109  119  129  139  152  178  222  232  242  27   37   47   57   67   77   87   97
10   11   12   13   14   153  18   223  233  243  28   38   48   58   68   78   88   98
100  110  120  130  140  155  19   224  234  244  29   39   49   59   69   79   89   99
101  111  121  131  144  16   2    225  235  246  3    4    5    6    7    8    9
102  112  122  132  145  160  20   226  236  249  30   40   50   60   70   80   90
103  113  123  133  146  17   200  227  237  25   31   41   51   61   71   81   91
104  114  124  134  148  170  204  228  238  250  32   42   52   62   72   82   92
105  115  125  135  149  171  205  229  239  251  33   43   53   63   73   83   93
106  116  126  136  15   175  206  23   24   252  34   44   54   64   74   84   94
107  117  127  137  150  176  21   230  240  253  35   45   55   65   75   85   95
108  118  128  138  151  177  22   231  241  26   36   46   56   66   76   86   96
-bash-3.00$ cd 11
-bash-3.00$ ls
11               11.SSschema      11.d11           11.plt002.exp    11.plt016.exp
11.DFedits       11.SSvars        11.d12           11.plt003.exp    11.plt017.exp
11.DFstudyPerms  11.SSvisitmap    11.d13           11.plt004.exp    11.plt018.exp
11.ICcenters     11.d02           11.d14           11.plt005.exp    11.plt019.exp
11.ICimages      11.d03           11.d15           11.plt006.exp    11.plt020.exp
11.ICkeys        11.d04           11.d16           11.plt007.exp    11.plt021.exp
11.ICqcs         11.d05           11.d17           11.plt008.exp    11.plt086.exp
11.ICrecords     11.d06           11.d18           11.plt009.exp    11.plt501.exp
11.ICschema      11.d07           11.d19           11.plt010.exp    11.sas
11.ICvisitdates  11.d08           11.d20           11.plt011.exp
11.ICvisitmap    11.d09           11.d23           11.plt012.exp
11.SScenters     11.d10           11.plt001.exp    11.plt013.exp
-bash-3.00$ pwd
/ous2/migration/logs/SNAP401/11
```

## Compare Them

- dircmp -d SNAP383 SNAP401 > snap383-to-snap401.out
- We leave off the –s because we want verbal confirmation of identical files
- Of course time/date stamps on report headers will be different
- Exports should be identical
- Report contents should be identical with exception of timestamps

## What can go wrong

- Export data is different
  - Index files are not resorted prior to data copy
  - Validation level from meta versus from raw data
- Report content is different
  - DF_SSschema shows choices ranges “1,2,3” in old versus “1,2,3,…” in new reports
  - DF_ICimages showing some differences in deleted images
  - Report formatting might have changed
- Account for the differences (expected versus unexpected differences)
- Document, document, document

## Lessons Learned

- Ensure memory mapped index files are not changing during migration
  - DataFax is live for exports/reports so a resort could occur and affect pre and post exports/reports
  - The modification time on index files doesn't change as they are memory mapped. So caution using "rsync" or "find" for copying.
- Keep logs of all activities for post migration review
- Ensure results of copies are checked for any issues

## Final Check

- Scripted data and report checking is not enough
- PHRI created over 700 validation test cases in addition to the ATK to ensure DataFax meets our intended use. These are run in a test environment prior to migration being approved.
- A subset of those 700 tests are run as sanity checks after migration and prior to production release to ensure the functionality of the system is as expected

## Questions and Answers

- Thank you