Data Warehouse

Moving forward planning – October, 2018

Near-Term Next Steps:

- ○ Nov/Dec 2018: data quality and client count conversation with DHS LT to introduce initiative, agree on key data elements (e.g., name, DOB, gender, language, race, and ethnicity), and obtain Division/system contact persons
- ○ January 2019: begin testing R algorithm version 2, reaching out to SDAL as needed
- ○ January 2019: ISB and DHS QA institute quarterly process to enhance data quality for key fields as identified by LT:
  - § ISB and Division/system contact persons standardize data entry formats
  - § ISB generates and Division/system contact persons review quarterly reports on missing data and potential duplicate clients; Division/system contacts work with staff to address
  - § ISB generates and Division/system contact persons review quarterly client counts from Data Warehouse for reasonability
  - § ISB updates system documentation to reflect changes

Commented [CG1]: This may not be practical, addition more data sources to algorithm is already a sticking point. Martha…clarify

Commented [MC2R2]: Reviewing or testing version 2 is not a good use of technical resources at this point. We are having issues adding systems to the modified version 1. The way R algorithm is architected is not scalable; it requires recoding of the final step 4 (Deduplication across all systems). Aaron is taking a look at the problem and will get back to us once he has a solution.

Commented [CG3]: Martha: please review and comment

Commented [MC4R4]: January 2019 is still doable to start generating quality reports on source systems;; based on the agreed data elements on the first bullet. It will also be possible to generate client counts by programs within each system for validation.

Coordinating the various topics from the review meeting last week with SDAL

1. Decided?
   a. Drop the old DW processes and counting.
   b. "Restating" the DHS unduplicated client count: seemingly deferred, with an informally stated need to for a summary available as needed (e.g., a "one page release" for future reference).
   c. Discussion of a new, additional "gold standard" client registry system was inconclusive.
2. Shift in warehouse record keeping
   a. Relate counts to the programs as identified in the budget.
   b. Define how programs not in the budget, but with data in the warehouse will be rolled up to the ones that are
   c. Focus on client counts – participation – in a fiscal year/FY or on a user defined reporting timeframe. This allows a shift from attempting monthly participation which is problematic for enrollment based program areas (e.g., SNAP).
   d. Shift to encounter based (services delivered) counting to determine active program participation during a reporting cycle.
   e. Client counts and program active participation will be reported in ranges (e.g. 37,000-38,000 Active DHS clients in fyxx).
   f. Reports generated from data warehouse are to be used for analysis, strategy, forecasting (e.g. budget preparation, new service initiatives) not to replace reporting from source systems (e.g., budget narratives, performance measures).

Commented [MI5]: Dropping entirely – or keeping as one potential analysis layer?

Commented [CG6R6]: Good point. Martha: Can we keep the old algorithm but for use on the new method for ingesting data? If so, this should be revised.

Commented [MC7R6]: Yes, we are going to keep the old algoritm, we will have to use the new method of data pull and modify how client program enrollment is counted. That will not only be a potential analysis layer; but an insurance policy for us in the "R" intitiave doesn't work for us at the end

Commented [CG8]: This was a key point…I thought…is it simply not relevant? That participation means "were enrolled at some point during" or "was an event during"?

3. Priorities
    a. Where could additional (contract) resources be the most valuable?
    b. Ideas (below):
        i. #1?…but need a charter first
        ii. #3 focus for ISB, technical but also with a strategic view on data warehousing
        iii. #4 near term, potentially simpler to find, tech focus (skills in R). May not be best long term investment.

Organizing the data warehouse efforts

1. Data Governance
    a. Ownership, expectations, communication plan – charter needed.

2. "Ingest" – the process of loading data from source systems into the data warehouse, including some mapping of fields.
    a. Current systems… Webvision, ETO, Cerner, VH-HCV, VH-LC&B, VaCMS
    b. Planned systems… Peerplace (including ASAP), DMC, ETO-HMIS,
    c. Possible systems . . . Real Estate Tax Relief, Adult Day Care
    d. Inaccessible systems, for which alternative data source will need to be maintained and provided by program staff… OASIS, WIC, SYNERGY (School Health), Community Corrections,

3. Program based (consumers of the data)
    a. Ingest vs. data warehouse standards (e.g., keep every nightly load or archive snap shots of monthly, FY(s)?
    b. Publishing a consistent, reliable, understood set of source system data (which is not all source system data) It should be a broader set of data (meaning more than 1 system; otherwise it will have limited value since data is available in the system of records.
    c. Mapping to budget programs from source system records.
    d. Utilizing visualization tools more widely in DHS (e.g., PowerBI or Tableau)

4. Unduplicated client counts – cross program statistics
    a. Currently dependent on the algorithm implemented in 'R', which may not scale well to larger number of source systems. Version 1 (currently in use), version 2 (untested)
    b. Includes creating a "statistical client" (if you merge several client records).
    c. Will always be an approximation; does not generate a "master list of clients".

5. Population based statistics
    a. Utilize #4 to develop DHS wide population statistics

6. Customer Service Tool
    a. Utilize #3 to provide a broader view to staff of how a client relates to DHS and how they are currently being served.
    b. Utilize #4 to aid staff in identifying client records across multiple systems

Commented [MI9]: Would contract resources help here?

Commented [CG10R10]: True…personal preference to have at least a draft charter to help focus the recruiting.

Commented [MI11]: I think we're intending to begin working with what we have, while simultaneously addressing data governance from multiple angles. Would it be possible to remove or reframe Data Governance here? Listing it first makes it look like a prerequisite, which would delay progress for potentially a long time frame…

Commented [CG12R12]: Uh…perhaps less informal, but looking at the bullets you drafted above assuming roles and monthly activities starting January, I think we need something…?

Commented [MI13]: Just confirming – is this already included as part of ETO, or will HMIS be separate?

Commented [MI14]: I believe this is going to be in PeerPlace

Commented [MI15]: I added this comment, because I thought it was important to capture Anita's intention that the data warehouse eventually include data from all programs – is there a better way/place to reflect this?

Commented [CG16R16]: Yes, lets. I listed "inaccessible" not to indicate 'never', but 'not at the moment'. So, naturally, I reacted to your insertion as "do these too…now"!

Commented [MI17]: School Health let me know that they have the data now to derive unduplicated client counts from SYNERGY, so perhaps move this to "Possible Systems"

Commented [CG18R18]: Excellent, would be valuable.

Commented [CG19]: I may be missing something…is this text adding some special consideration that isn't implied throughout?

Commented [MI20]: I thought it was intended to be scalable across as many systems as we need?

Commented [CG21R21]: Martha can clarify…