

REGIONAL & STATEWIDE DATA WAREHOUSING

1

Data Warehouse System Design and Technology Choices: Regional & Statewide HMIS/Human Services Projects An advanced curriculum



This curriculum was prepared by the Cloudburst Group under cooperative agreement MDMV00107 with the Department of Housing and Urban Development's (HUD's) Office of Community Planning and Development. This curricula was developed by Ray Allen, Tony Gardner and John Bardos under contract with the Cloudburst Group.

LEARNING OBJECTIVES

2

- To introduce the audience to:
 - System design and technology challenges and choices that arise in the development of HMIS/human services data warehouse projects;
 - The solutions developed and critical lessons learned in existing data warehouse projects;
 - The challenges, solutions and lessons of two HMIS/human service data warehouse projects--the San Francisco Bay Area RHINo and Michigan State SHADoW projects.

TRAINING OUTLINE

3

- Data Integration
- RHINo and SHADoW Project Background
- Program Goals and Requirements
- Architecture and Systems Overview
- General Design Considerations
- Data Specifications
- Data Processing
- Systems and Network Structure
- Data Warehouse Reporting Considerations
- Database, ETL, and Analysis Software Considerations
- Confidentiality and Security
- Summary of Significant Challenges
- Summary of Lessons Learned

DATA INTEGRATION

4

- Data Integration is combining data residing in different sources and providing users with a unified view of these data (Wikipedia).
- Data Integration or merging of data can take several routes:
 - **XML Data Sharing** – sharing a common case file for clients (or other data field)
 - Example – Michigan’s Muskegon project
 - **Combining Systems** – combining several similar systems (e.g. HMIS) into a single system
 - Example – 9 CoCs in Louisiana in a single HMIS
 - **Data Warehousing** – extracting, transforming and loading (ETL) data from several sources into a single queryable schema
 - Examples – San Francisco Bay Area’s RHINo and Michigan’s SHADoW project

RHINo AND SHADoW PROJECT BACKGROUND

5

- The San Francisco Bay Area RHINo Project and State of Michigan SHADoW Project are existing data warehouse projects.
- They are used throughout as examples only to illustrate challenges, solutions, and lessons learned in HMIS/human services data warehouse design.
- A brief review of the background of each project will help put the system design and technology choice issues into context.

REGIONAL DATA WAREHOUSE EXAMPLE: BAY AREA RHIN₀ PROJECT

6



- Geography: San Francisco and Monterey Bay Area
- Source Data: HMIS data from 11 Counties/CoCs
- Objective: Provide a rich repository of regional data to better analyze trends, gaps in services, and mobility patterns among homeless people, and to inform regional policy and funding directions.
- Key Planning Group: Bay Area Counties Homeless Counties Information Collaborative

RHINo and SHADoW PROJECT

BACKGROUND - RHINo

7

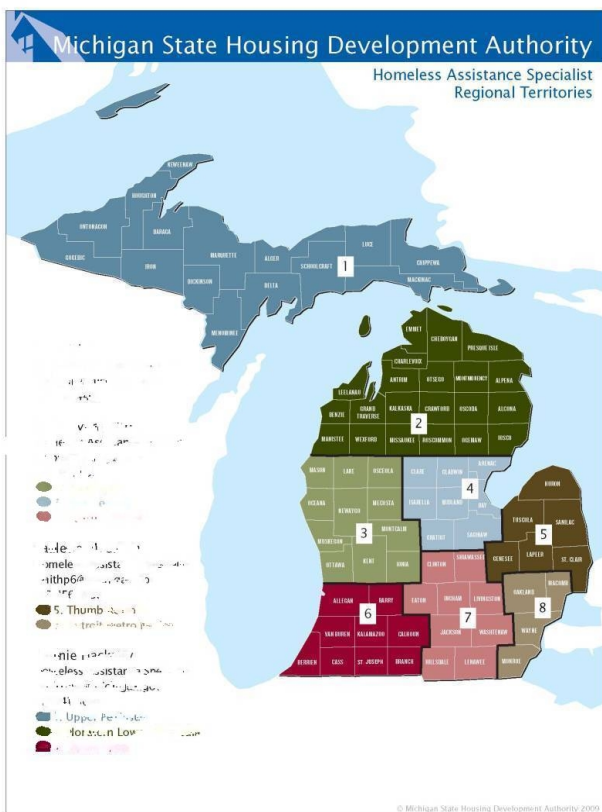
- **Bay Area Counties Homeless Information Collaborative**
 - **Informal Collaborative:** Eleven Continuums of Care in the San Francisco and Monterey Bay Area meet to discuss regional issues around homelessness.
 - **Mission:** To better enable policy makers, service agencies, and funders to understand and service the needs of the homeless within the community
 - **Goals:**
 - ✦ Obtain an unduplicated regional count of homeless persons
 - ✦ Identify prevalence of cross-county chronic homelessness
 - ✦ Understand client movement across continuum boundaries
 - ✦ Analyze service usage across continuums
 - ✦ Inform funders about effectiveness of sponsored programs in the region

RHINo and SHADoW PROJECT

BACKGROUND - SHADoW

8

- Source Data: Statewide HMIS and Michigan Human Services Data Warehouse
- Objective: Examine use of state mainstream systems to help determine: cost of homelessness, impact of state program changes, patterns of state service usage relating to homelessness, and extent homeless are benefiting from state services.
- Key Planning Group: SHADoW Leadership Board



RHINo and SHADoW PROJECT

BACKGROUND - SHADoW

9

- SHADoW
 - Mission: Understanding the needs of homeless households and promoting greater responsiveness, accountability, and impact of public and private homeless services.
 - Goals:
 - ✦ Understand the cost of homelessness to Michigan State's systems of care.
 - ✦ Track the impact of changes in state programs and allocations on the numbers of homeless, the characteristics of those served, and the effectiveness of services in reducing homelessness.
 - ✦ Explore patterns of service usage (both state and private) that relate to patterns of homelessness.
 - ✦ Determine if homeless persons are benefiting from state services designed to help them.

PROGRAM GOALS AND REQUIREMENTS

10

- Once Data Warehouse option has been chosen:
 - Data warehouse program goals and requirements must be clearly defined BEFORE any systems design and technology choices are made.
 - To do otherwise risks misunderstandings, planning confusion and delays, and flawed system that does not meet its intended purpose.
 - Remember – the program governs technology and not the other way around!

PROGRAM GOALS AND REQUIREMENTS

11

- Start by asking:
 - What is the purpose of the system?
 - What **must** it be able to accomplish?
 - How will success be defined?
 - Who will contribute data?
 - What kinds of reports or other data outputs will be needed to accomplish the project's goals?
 - What specific data elements will be needed to produce the desired reports or other data outputs?
 - What level of data security and confidentiality will be needed?
 - Will personal client information be included?

ARCHITECTURE AND SYSTEMS OVERVIEW

12

- A data warehouse merges large amounts of data from multiple sources and makes them easy to retrieve and use in reports.
- The nature of the design challenge is to structure the overall data warehouse system in a way that meets the needs of the participants, including those who:
 - Provide the data
 - Analyze the data
 - Manage the data warehouse
- Ensure the design takes account of the needs of:
 - Clients
 - Policy makers
 - Service Providers

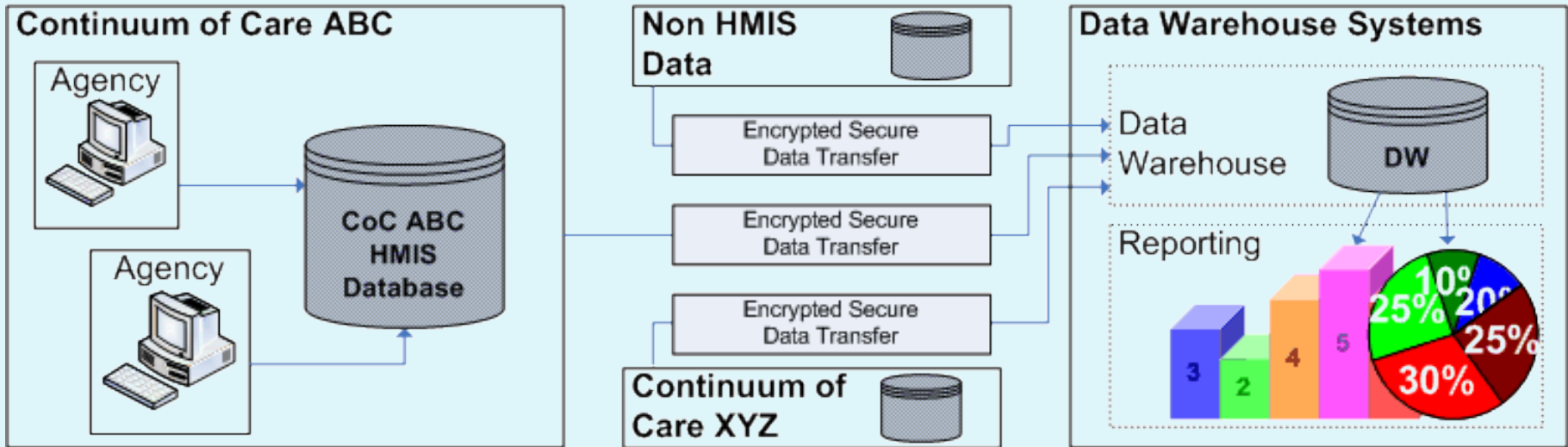
ARCHITECTURE AND SYSTEMS OVERVIEW

13

- Do not make the system too complex or costly for the participants.
- The data that will eventually populate the warehouse is collected in different systems. These systems may use different collection strategies, processes and definitions.
- Each source converts its data to the export data specifications.
- Data is encrypted at the source, and is securely transferred to the data warehouse systems.
- The data warehouse systems transform source data into the warehouse.
- Reporting and analysis occur using the data warehouse.

ARCHITECTURE AND SYSTEMS OVERVIEW

14



GENERAL DESIGN CONSIDERATIONS

15

- Start with the end in mind: Define the reporting and analysis requirements. They will determine which source data are required.
- Obtain buy-in from partners on type and format of reports needed.
- Review data standards for the industry and source data. HUD provides HMIS Data Standards.
- Review each of the source systems and any existing data standards, together with the requirements, and define export data specifications for the source systems.
- Identify gaps or variations between:
 - Data standards and data export specifications
 - Actual data being collected by each source system

GENERAL DESIGN CONSIDERATIONS

16

- Understand how source data are stored and updated in the source system and what is required to properly convey data changes to the data warehouse.
- Determine how often changes in the source system need to be reflected in the data warehouse – daily, monthly, quarterly?
- Determine data transfer and disk space requirements - plan for growth.
- Consider data, network, and server security requirements.
- Define report and data access security requirements for *end users* to access the data warehouse reporting system, and generated reports.

DATA SPECIFICATIONS

17

- The challenge
 - Most source system data differs from one another. How do you merge data when the data standards, data elements, allowable values, and data entry rules all differ?
- The solution
 - Identify a common data specification (or “file format” or “schema”) for standardizing and exporting data.

DATA SPECIFICATIONS

18

Creating a common data specification across multiple source systems – and determining how to handle differences between source systems and Null data – is a critical aspect of building a data warehouse.

- Export data specification

- Each source system will convert and export its data to a common export data specification.
- HMIS may use the HUD HMIS CSV or XML data export specifications, both available for download at:
<http://www.hmis.info/>.
- Non-HMIS will have separate export data specifications, however, common elements between HMIS and non HMIS systems should use the same data specifications.

DATA SPECIFICATIONS

19

- **Rigorous data validation**
 - The source system must validate that all exported data satisfies the export data specifications
 - This same data validation will occur in a stage area of data warehouse, prior to transforming and loading into the warehouse.
 - Plan data specifications, data warehouse and reporting specifications to allow for Null values.

DATA SPECIFICATIONS: CSV & XML

20

- CSV and XML can generally represent the same data. The HUD CSV and XML data specifications are very similar in that they contain the same source data definitions.
- In a CSV file, different data elements are separated by a comma, and different sets of data (Client, Program, etc) are stored in different files.
- In an XML file, different data elements are surrounded by *tags* – `<LegalFirstName>Jane</LegalFirstName>` and typically all of the data (Client, Program, etc) are stored in one file.
- CSV is generally considered easier to work with, and most systems provide a way to easily export and import CSV data.

DATA SPECIFICATIONS: CSV & XML

21

- XML is a more rigorous file format specification, and there are many tools available that make it easy to read and write XML formatted data.
- XML can include an XML Schema Definition (XSD) companion file that specifies the exact data specification and format for an XML file. Using an XML tool, the XSD can be used to validate the source XML data file, and this validation can occur when the XML file is created at the source, and when the XML data file is imported by another system.
- Validation of CSV data can be similarly done using database constraints by both the source system and the importing system.
- With either file format, the same validation must be defined and implemented by all systems using the data.

DATA SPECIFICATIONS: RHINO EXAMPLE

22

- RHINO uses only HMIS data so it followed the HUD HMIS Data Standards.
- RHINO selected the HUD CSV rather than HUD XML export specification because it was easier and cheaper for source HMIS staff to use.
- The HUD CSV format was modified and extended into what is called the RHINO CSV File Format Specification.
- RHINO participants reviewed their data for conformity to the HUD Data Standards and the RHINO CSV File Format Specification.

DATA SPECIFICATIONS: RHIN_o EXAMPLE

23

- When exceptions were found to exceed a small percentage of the total data collected, then that exception was considered for inclusion in the RHIN_o export specification
- RHIN_o added only a few data extensions to the HUD format:
 - Race: “Other”
 - Gender: “Transgender”
 - An “anonymous” designation for clients
 - The zip code of the agency providing the service to the client

DATA SPECIFICATIONS: SHADoW EXAMPLE

24

- SHADoW includes data from diverse systems including: Michigan's Statewide HMIS, Department of Human Services (multiple databases) and Department of Community Health Medicaid database.
- Data is transformed within the data warehouse from diverse sources using Michigan States UCI (Unified Client Index).

DATA SPECIFICATIONS: SHADoW EXAMPLE

25

- The transformation yields the following simple data structure:
 - Person Characteristics
 - Event Information
 - ✦ Age of Consumer
 - ✦ Date of Event
 - ✦ Event Type
 - ✦ Cost of Event (as appropriate)
- Event Types reflect transactions within the source systems such as emergency shelter or transitional housing intake (HMIS), medicaid payment (DCH), “SER Rental Assistance,” foster care intake, ect.

DATA PROCESSING

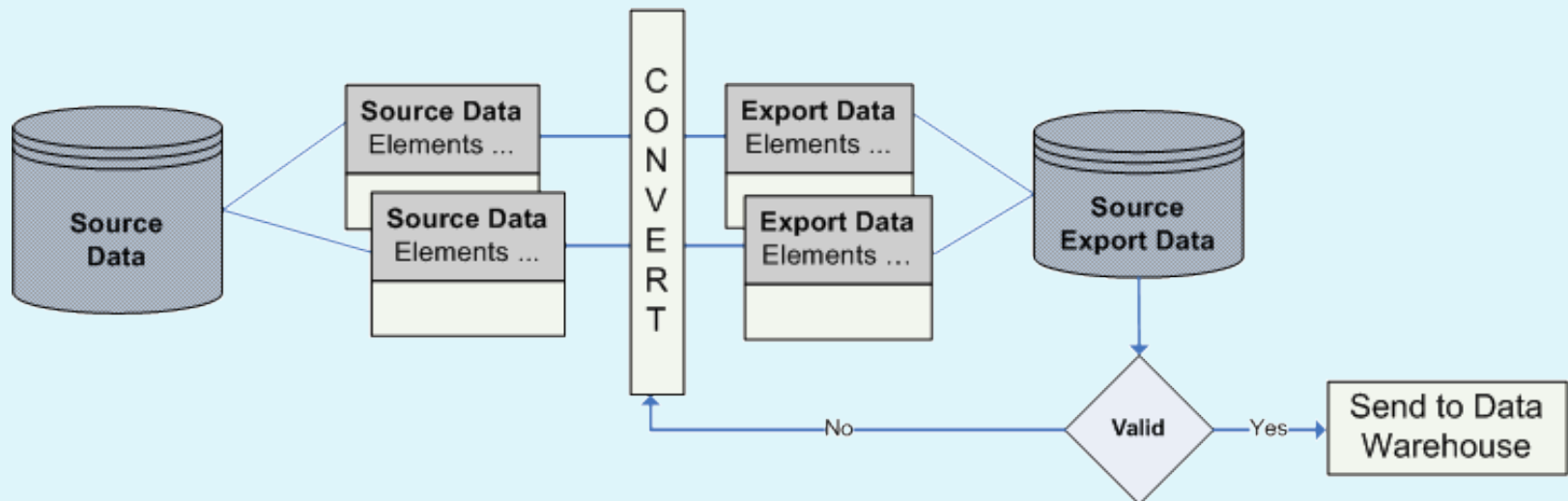
26

- **The challenge**
 - How do you take data from the source systems, prepare them, and move them to the data warehouse?
- **The solution**
 - Develop and implement an effective “Extract, Transform, and Load (ETL) process.”

DATA PROCESSING - DATA EXTRACTION

27

- Source data are converted at the source into the export data specification.
- Source export data are validated at the source.
- Source data that do not satisfy the export data specification must be corrected by the source system.



DATA PROCESSING - DATA EXTRACTION

28

- Data warehousing is often interested in changes to source data over time.
- A source system may only store the current data for a client, and may not store each change to each data element (history) over time. In this case, it may be necessary to extract data daily, so that changes over time will be accumulated in the data warehouse.
- HMIS systems typically will store the history to changes in client data. Consequently, data can be extracted at *any* time interval and can contain the full history of changes.

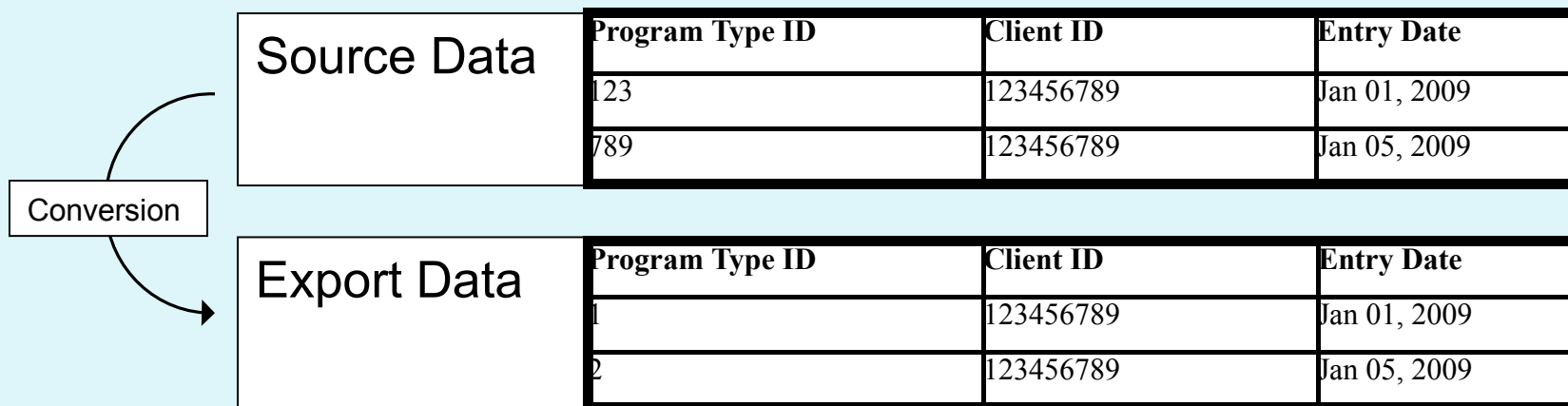
DATA PROCESSING - DATA EXTRACTION

- Source systems can either extract all of their data each time it sends data to the data warehouse, or extract only the data that has changed since the last time it sent data to the data warehouse.
- Sending all of the data is typically easier to develop, however each export data set may be large, and will continue to increase in size over time.
- Sending only the changed data minimizes the load and impact on both the source system and the data warehouse system, and minimizes the growth in the size of export data over time. However, this approach may require additional development costs.

DATA PROCESSING - DATA MAPPING

30

- Source data typically will have their own specific *IDs* to uniquely identify individual data, and types of data. These IDs are often automatically generated by the database and are hidden from end users.
- For example, a database may have the ID “123” for an Emergency Shelter type program.
- HUD Data Standards have the ID “1” to identify an Emergency Shelter program.
- Data being transferred will need to be mapped and converted from the source database specific IDs to the export data specification IDs. For example:



DATA PROCESSING - DE-IDENTIFIED DATA

31

- The source data likely contains personal protected information (PPI), including:
 - Full Name
 - Social Security Number
 - Date of Birth
- It may not be acceptable to transfer PPI to the data warehouse, however it is valuable to identify the same person across multiple database sources.

DATA PROCESSING - DE-IDENTIFIED DATA

32

- Define a **UID** that will uniquely identify the same person across different source systems.
- Create the **UID** in each source system and add to the export data.
- The **UID** should not contain any complete PPI.
- Remove PPI from source data prior to transfer to the data warehouse.

DATA PROCESSING: DATA TRANSFER

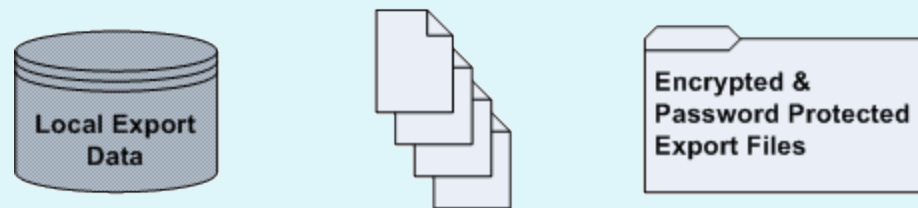
33

- Source data needs to be transferred to the data warehouse.
- Two general ways to transfer data:
 - Transfer data directly from one database to another database
 - Extract data to flat files (CSV or XML), then transfer the files
- Direct connection – The source database will make a direct connection to a data warehouse database, and will directly transfer data from one database to another. Data must be encrypted when transferred through a network, such as an encrypted virtual private network (VPN).
- Flat files – Source data is extracted into flat files that must be encrypted and then transferred using technologies such as Secure Shell (SSH) and Secure Copy (SCP).

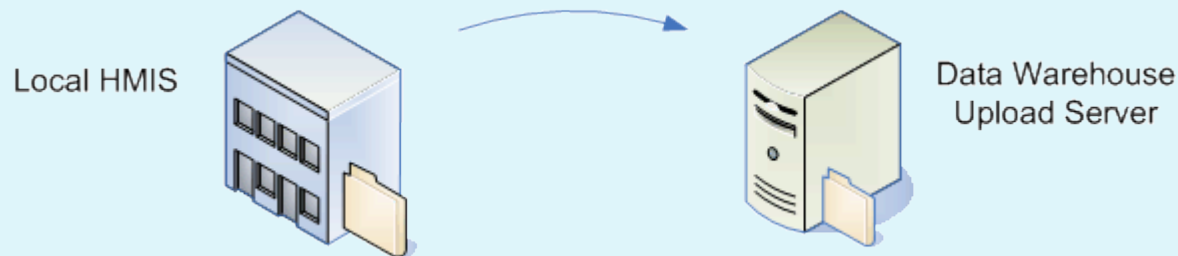
DATA PROCESSING: DATA TRANSFER – EXAMPLE

34

- Source database data is extracted out of the database and converted to the export data specification, and validated to the export data specification.
- Export files are combined into an encrypted file.



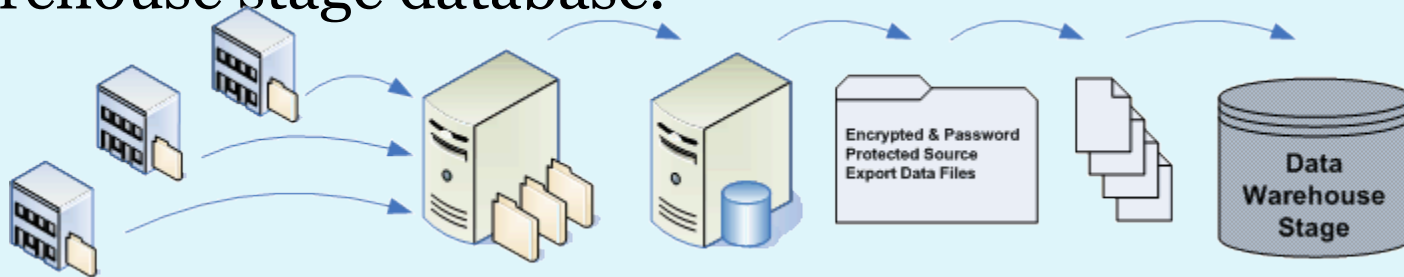
- The encrypted file is transferred from each source via an SSH secured network to the data warehouse server.



DATA PROCESSING: DATA TRANSFER – EXAMPLE

35

- Encrypted source files from multiple CoC's are automatically decrypted and loaded into the data warehouse stage database.



- Data is validated to the export data specification, and data not conforming is not loaded. The data is transformed into the data warehouse for reporting and analysis.



DATA PROCESSING - DATA WAREHOUSE TRANSFORMATION AND LOADING

36

OLTP – Online Transaction Processing

- Source systems use an OLTP type database schema in which the data is broken down into a large number of separate tables.
- Additionally, the data is stored in a *normalized* schema, in which a given data value will be stored only once in the database.
- These types of databases are designed to support large volumes of transactions – a lot of small changes to the data occurring frequently throughout the day.
- However, because the data is broken down into a large number of tables, it can be slower to query large amounts of data for reporting and analysis.

DATA PROCESSING - DATA WAREHOUSE TRANSFORMATION AND LOADING

37

OLAP – **O**nline **A**nalytical **P**rocessing

- Data warehouses use an OLAP type database schema in which the many separate tables of source data are combined into a few tables of data warehouse data.
- This type of schema is designed for reporting and analysis rather than transaction processing.
- This may cause the same data to be repeated in the database, but this will improve the reporting and analysis performance.

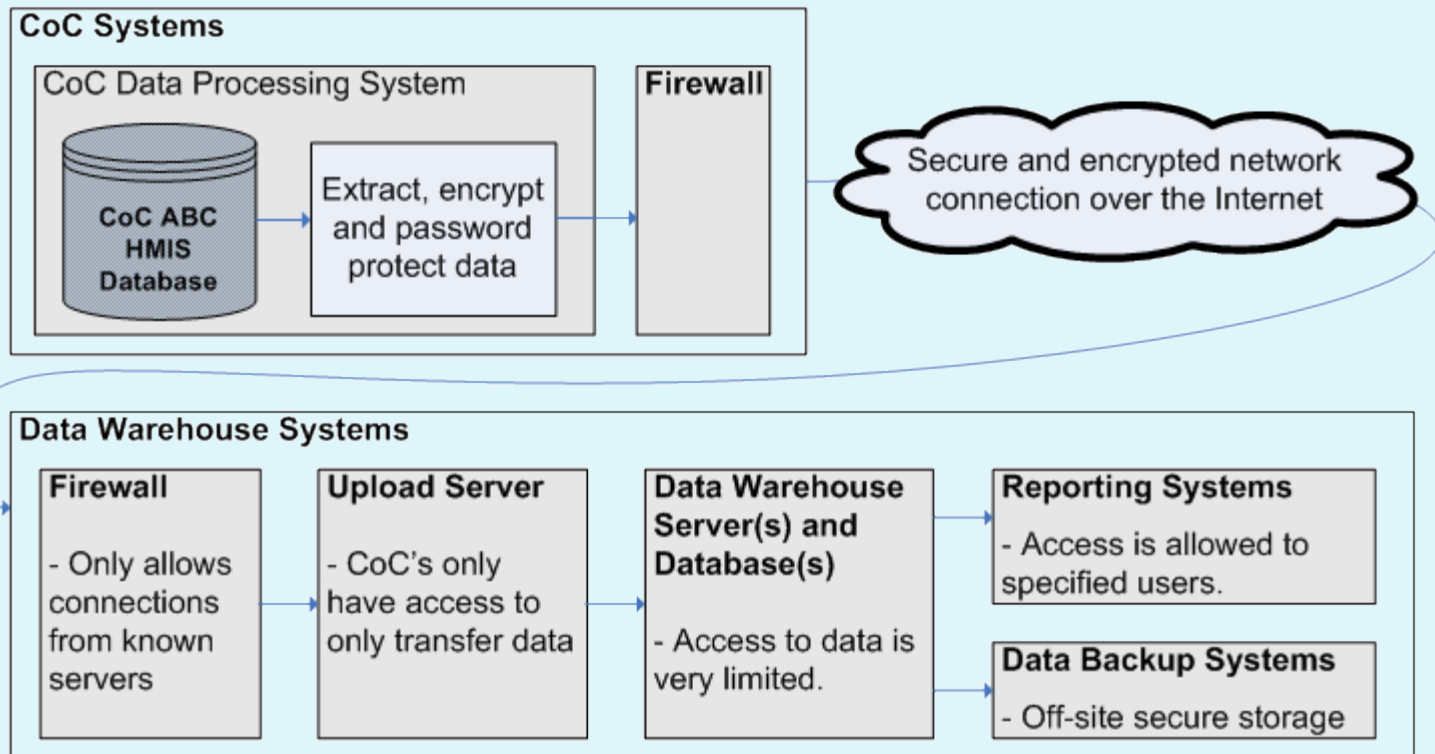
DATA PROCESSING - DATA WAREHOUSE TRANSFORMATION AND LOADING

38

- During the data warehouse transformation and loading, summary and analysis calculations may be computed and stored.
 - ✦ For example, a data warehouse can pre-calculate and store the number of members in a household. This calculation is done once when the data is transformed and loaded into the data warehouse.
- Other transformations can include changing some source data to be consistent for all sources.
 - ✦ For example, if one source uses the two letter abbreviation for States, and another source uses the full spelling, the transformation process will convert the source data once during loading so that it is the same for all sources in the data warehouse.

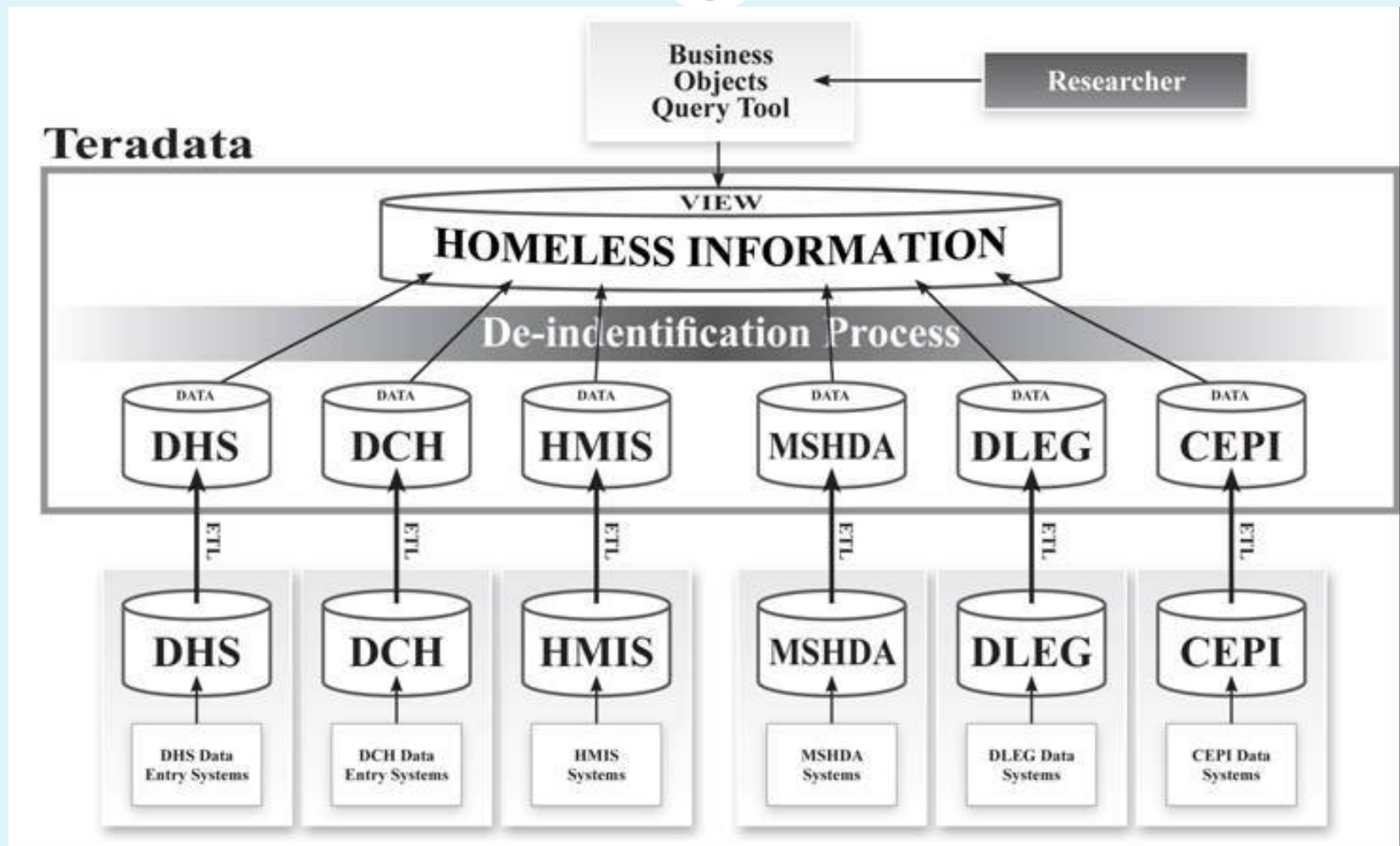
SYSTEMS AND NETWORK STRUCTURE - RHINO EXAMPLE

39



SYSTEMS AND NETWORK STRUCTURE - SHADoW EXAMPLE

40



DATABASE, ETL, AND ANALYSIS SOFTWARE CONSIDERATIONS

41

- Choosing the data warehouse database - open source or proprietary?
 - Leading open source systems are stable, reliable, and generally free.
 - Proprietary systems are well supported, with managed support and software upgrade processes.
- The ETL code can be developed using tools specific to data warehouse ETL, or can be custom developed.
- There are both open source and proprietary ETL technologies.

DATABASE, ETL, AND ANALYSIS SOFTWARE CONSIDERATIONS

42

- Data warehouse ETL technologies may have a learning curve, but provide better documentation for the data transformations, and may make it easier to assess and implement changes.
- Reporting and analysis can be developed using open source or proprietary systems. Some database vendors may also provide reporting and analysis systems.

CONFIDENTIALITY AND SECURITY

43

- The protection of client information should be paramount.
- If personally protected information (PPI) such as name, SSN, and date of birth are not being transferred to the data warehouse, then there has to be a mechanism for being able to de-duplicate the records from multiple sources.
- The data warehouse must have the **maximum** security in place to protect client data.
- Prepare to educate consumers from all aspects of the project regarding the uses of data as well as the de-identification protocols.

CONFIDENTIALITY AND SECURITY

44

- Security features should include:
 - Encryption during data transfer and storage
 - Firewalls
 - Anti-virus software
 - Anti-spyware software
 - Authentication and access controls
 - Segregated data upload and data analysis servers
 - Secure server hosting
 - Disaster and recovery services
 - Staff policies and procedures for data confidentiality
 - Audit trail and regular auditing

CONFIDENTIALITY AND SECURITY

45

- De-duplication of client data normally requires the use of PPI.
- Use of PPI could make the discovery of client theoretically possible.
- Remove PPI from data sets submitted to warehouse.
- Create a UID algorithm to uniquely identify the same client across exporting systems.
- Uses an algorithm to encrypt the UID.
- Generate the UID at each source system.
- Anonymous clients can not be de-duplicated.
- Test the de-identified data set by evaluating the possibility of re-identification using other available data sources.

SUMMARY OF SIGNIFICANT CHALLENGES

46

- Varying staff capacity and resources at the data sources.
- Varying levels of data quality and data collection standards.
- Lack of buy-in and consensus on key design issues.
- Lack of adequate funding.
- Difficulty in allocating resources and funding for extracting data from source systems.

SUMMARY OF SIGNIFICANT CHALLENGES

47

- Difficulty in affording technically qualified staff for building and managing the warehouse.
- Different perspectives on the “Ownership” of data.
- Concerns about client privacy and confidentiality.
- Concerns that comparative data will make a partner look bad.
- Different levels of readiness to move forward.
- ***People are more challenging than the technology!***

SUMMARY OF LESSONS LEARNED

48

- Reducing local costs and staff burden is the most valuable key for obtaining buy-in and consensus. The overall design must be simple and participation (after initial setup) fairly automatic.
- Some systems contributing to the data warehouse may have data that does not meet the export data specifications. Allow time for these organizations to update their source data to meet the specifications prior to transferring final data to the data warehouse.
- Get sample export data from source systems early in the process, and provide data validation and basic analysis results back to the sources. This will likely be an iterative process before “clean” final data can be provided.

SUMMARY OF LESSONS LEARNED

- Based on HUD Data Standards, some client *attributes* such as Veteran Status and Disabling Conditions are specified with Program Participation data. However, source systems may be collecting this data for clients that are not participating in a program. It is still valuable to transfer this data to the warehouse.
- You may find that the source data has many more NULL values than anticipated. While source systems may be working to reduce the occurrences of NULL values, you may need to plan your export data specifications, data warehouse, and reporting specifications to allow for NULL values as much as possible.

DATA WAREHOUSING: ADDITIONAL RESOURCES

50

- Curricula:
 - Data Warehousing 101
 - Data Warehouse Planning and Governance
 - Data Warehouse System Design and Technology Choices
- Documents:
 - What is a Data Warehouse?
 - What is SHADoW?
 - What are BACHIC & RHINo?
 - BACHIC Overview & Guiding Principles for RHINo
 - SHADoW Interagency Agreement
 - SHADoW Participation Agreement
 - SHADoW Data Use Agreement