

REGIONAL & STATEWIDE DATA WAREHOUSING

1

Data Warehousing 101: Regional & Statewide HMIS/Human Services Projects An introductory 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 Barb Ritter under contract with the Cloudburst Group.

LEARNING OBJECTIVES

2


- To introduce the audience to:
 - Basic purposes, concepts, and components of Homeless Management Information System (HMIS)/human services data warehousing;
 - The experience of two communities - Michigan and the San Francisco Bay Area - in developing HMIS/human services data warehouses; and
 - Key tips, solutions, and lessons learned in the areas of:
 - ✦ Planning processes and structures
 - ✦ Data warehouse design
 - ✦ Privacy and security



PACKET CONTENTS

3

- This training packet contains the following resources:
 - What is a Data Warehouse?
 - What is SHADoW?
 - What are BACHIC & RHINo?

- The  icon indicates a corresponding resource is available.

TRAINING OVERVIEW

4

- Data Integration
- What is an HMIS data warehouse?
- Why an HMIS data warehouse?
- HMIS data warehouse examples
- Planning and process requirements
- Programmatic requirements
- System design considerations
- Information process issues
- Privacy and security issues
- Significant challenges
- Key solutions and lessons learned

DATA INTEGRATION

5

- 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

WHAT IS AN HMIS DATA WAREHOUSE?

6



What is a
Data
Warehouse?

- A data warehouse is commonly understood to be a central database for organizing and analyzing data from more than one source.
- There are many possible ways to structure a data warehouse, such as several local HMIS databases feeding into a single warehouse, or a local HMIS database and a mainstream resource database feeding into a single warehouse.

WHAT IS AN HMIS DATA WAREHOUSE?

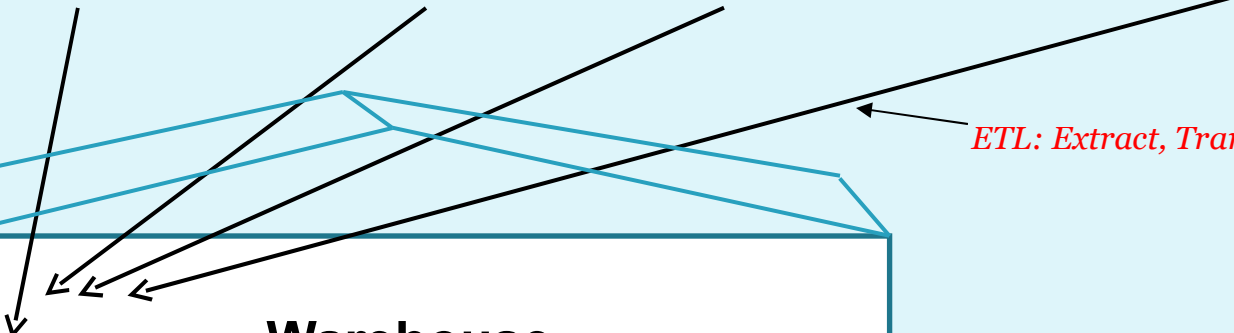
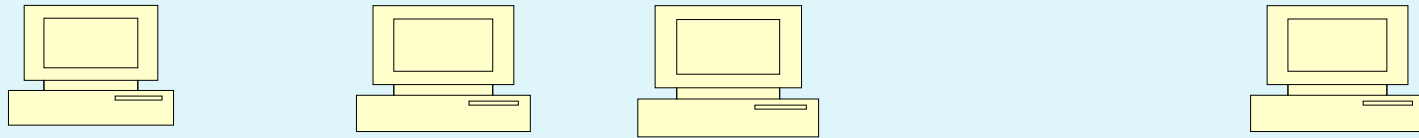
7

- Typical data warehouse components:
 - Data Source: Data elements to be collected from defined participants
 - Unique identifier: A way to de-duplicate records
 - Common “schema” and process for extraction, transformation, and loading of data (ETL)
 - A relational database
 - Security: Secure Socket Layer (SSL) connection, encryption, firewall
 - Analysis and reporting software

WHAT IS AN HMIS DATA WAREHOUSE?

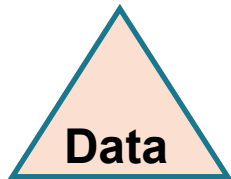
8

Data Sources (e.g., HMIS and/or state mainstream data systems)



ETL: Extract, Transform, Load

Warehouse



WHY AN HMIS DATA WAREHOUSE?

9

- A good data warehouse will:
 - Provide the RIGHT data.....
 - To the RIGHT people.....
 - At the RIGHT time.....
 - RIGHT NOW
- Data in – information out
- Multiple data sources can be analyzed in combination.
- Most reports require data from many tables.
 - Most HMIS are designed for data input and ease of day-to-day client transactions, but less so for ease of reporting
- Data warehouses are designed for ease of data retrieval, analysis, and reporting across large data sets.

WHY AN HMIS DATA WAREHOUSE?

10

- An effective tool for decision-making support
- The potential uses of HMIS/human data warehouses:
 - Analyzing regional or state demographics, trends, and outcomes
 - Assessing homeless use of mainstream services
 - Calculating the cost homelessness
 - Determining what interventions work to prevent and end homelessness
 - Informing the development of regional or state 10-year plans to end homelessness

WHY AN HMIS DATA WAREHOUSE?

11

- **Sample report: client movement between CoCs**

- **Description**

Count of clients who have received services
In at most *N-many* counties.

- **Data Elements**

Program Start Date, Program End Date,
Service Start Date, Service End Date,
Last Permanent Zip Code

- **Counties**

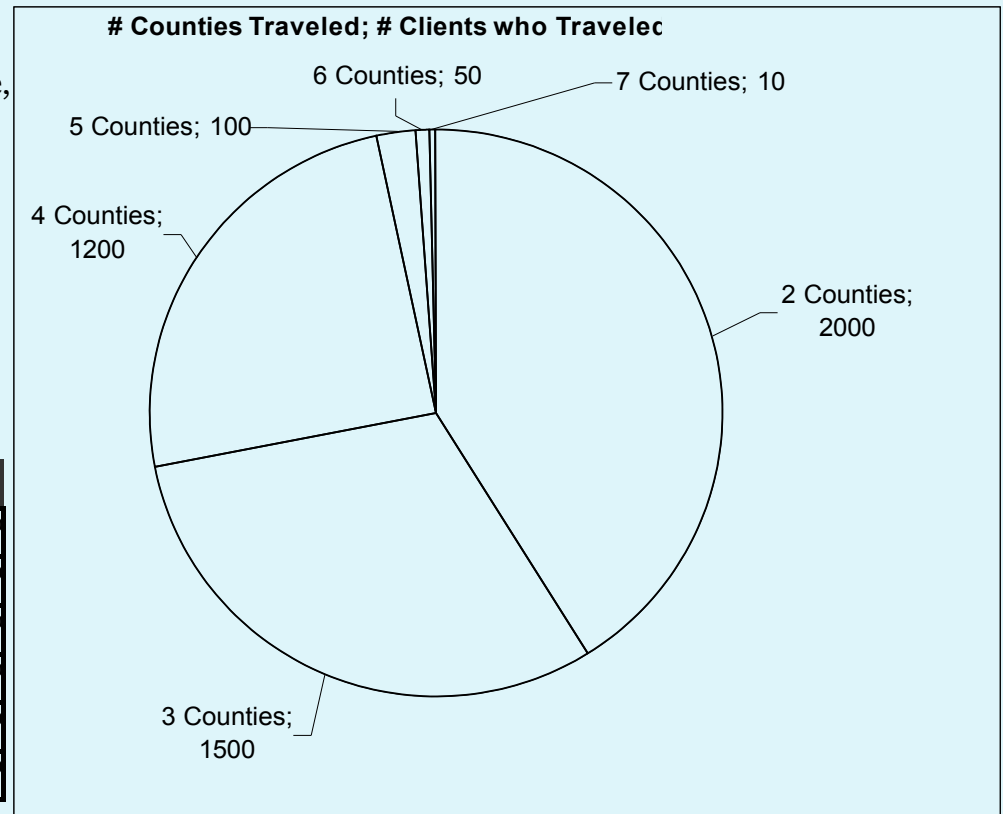
All Counties

- **Date Range**

2007

- **Data Statistics & Data Limitations**

# Counties Traveled	# Clients Traveled
2 Counties	2000
3 Counties	1500
4 Counties	1200
5 Counties	100
6 Counties	50
7 Counties	10



WHY AN HMIS DATA WAREHOUSE?

12

- Sample reports: Cost of homelessness combining state and HMIS data for a CoC.

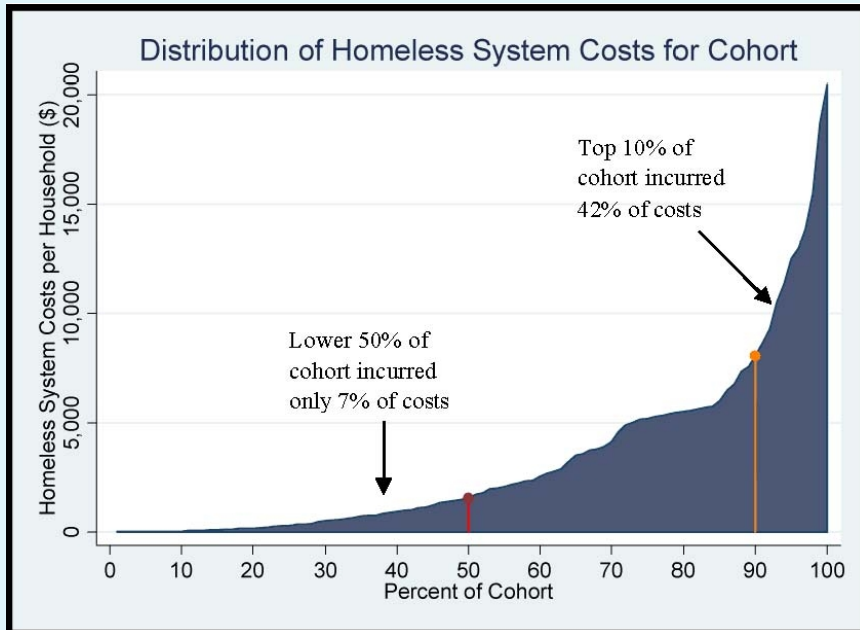
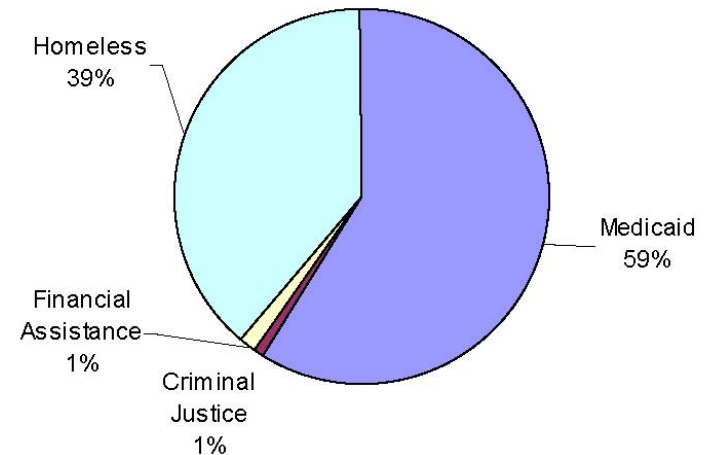


Exhibit 4: Distribution of Costs by Domain During Homeless Period



HMIS DATA WAREHOUSE EXAMPLES

13



What is SHADoW?
What are BACHIC &
RHINO?

- Michigan Statewide Homeless Assistance online Warehouse (SHADoW) Project
 - Contact: Barbara Ritter
Michigan Coalition Against Homelessness
15851 Old UU-27, Building 30, Suite 315
Lansing, MA 48906
- Bay Area Regional Homeless Information Network organization (RHINo) Project
 - Contact: Ray Allen
Community Technology Alliance
115 East Gish Rd., Suite 222
San Jose, CA 95112

STATEWIDE DATA WAREHOUSE EXAMPLE: MICHIGAN SHADoW PROJECT

14



- **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

STATEWIDE DATA WAREHOUSE EXAMPLE: MICHIGAN SHADoW PROJECT

15

- SHADoW background
 - Michigan is the 8th most populous state in the nation with a 2008 estimated census of 10,003,422 persons.
 - Over 500 social service and safety-net agencies from 82 of Michigan's 83 counties participate on a single HMIS platform - Michigan's Statewide HMIS (MSHMIS).
 - The Michigan Department of Information Technology operates a data warehouse that combines data from the various state departments. In building the warehouse, Michigan developed a Common Index that allowed records from diverse systems to be merged.
 - The SHADoW Project adapted the existing technology to be the first de-identified dataset strictly developed for the purposes of planning and research.

STATEWIDE DATA WAREHOUSE EXAMPLE: MICHIGAN SHADoW PROJECT

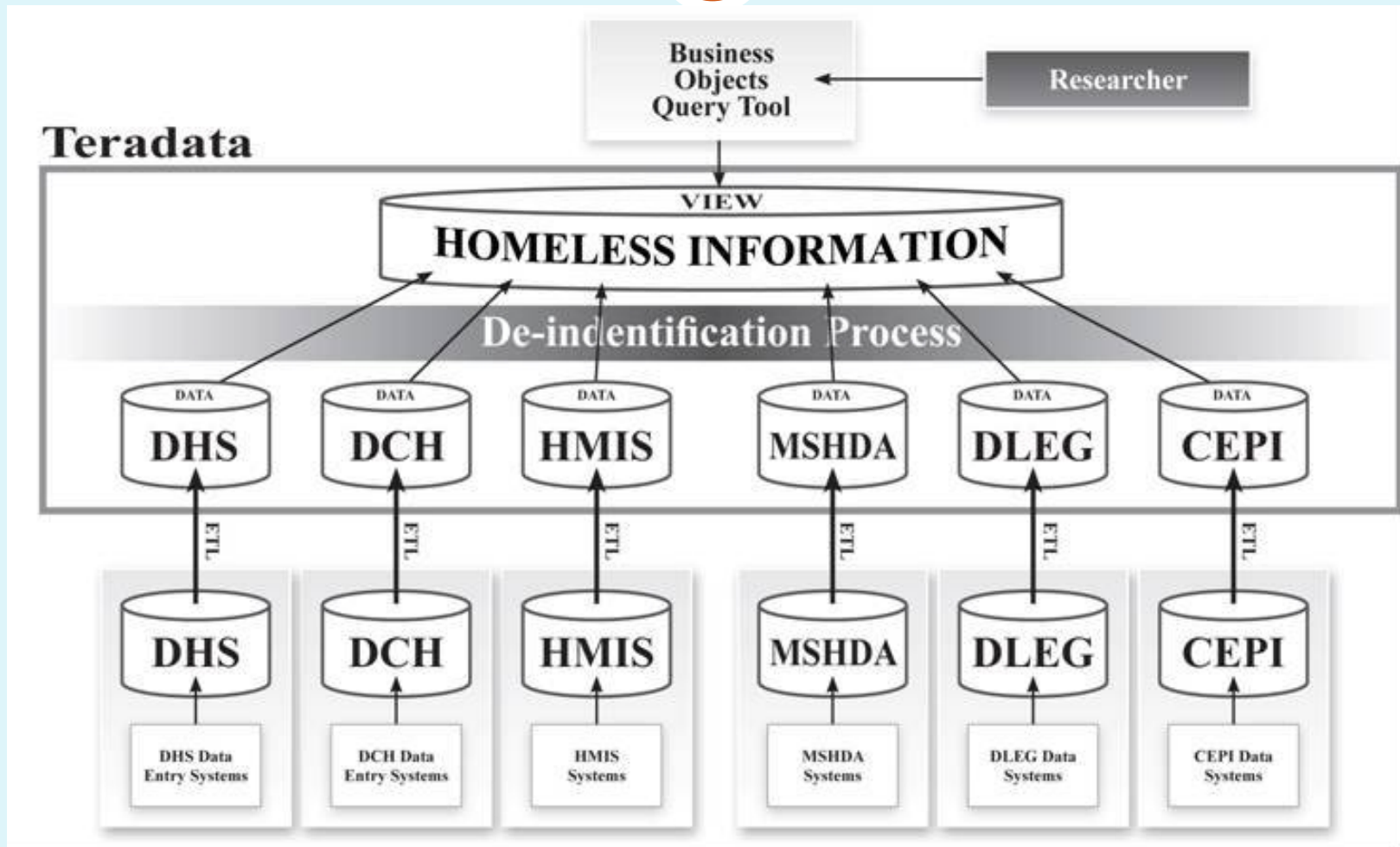
16

- 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 State of Michigan systems of care.
 - ✦ Track the impact of changes in state programs and allocations on the numbers of homeless, the characteristics of those served and 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.

STATEWIDE DATA WAREHOUSE EXAMPLE: MICHIGAN SHADoW PROJECT

17



REGIONAL DATA WAREHOUSE EXAMPLE: BAY AREA RHINo* PROJECT

18



- **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 Information Collaborative

*Regional Homeless Information Network (RHINo)

REGIONAL DATA WAREHOUSE EXAMPLE: BAY AREA RHIN_o PROJECT

19

- Bay Area background
 - The 11-county region of Northern California has a population of 7,617,857 and a geographic area of 10,691 square miles
 - The 11-county January 2008 point-in-time homeless count was 35,902 (known undercount)
 - Data from 11 CoCs, representing hundreds of service providers, is combined into a de-duplicated regional data warehouse

REGIONAL DATA WAREHOUSE EXAMPLE: BAY AREA RHINO PROJECT

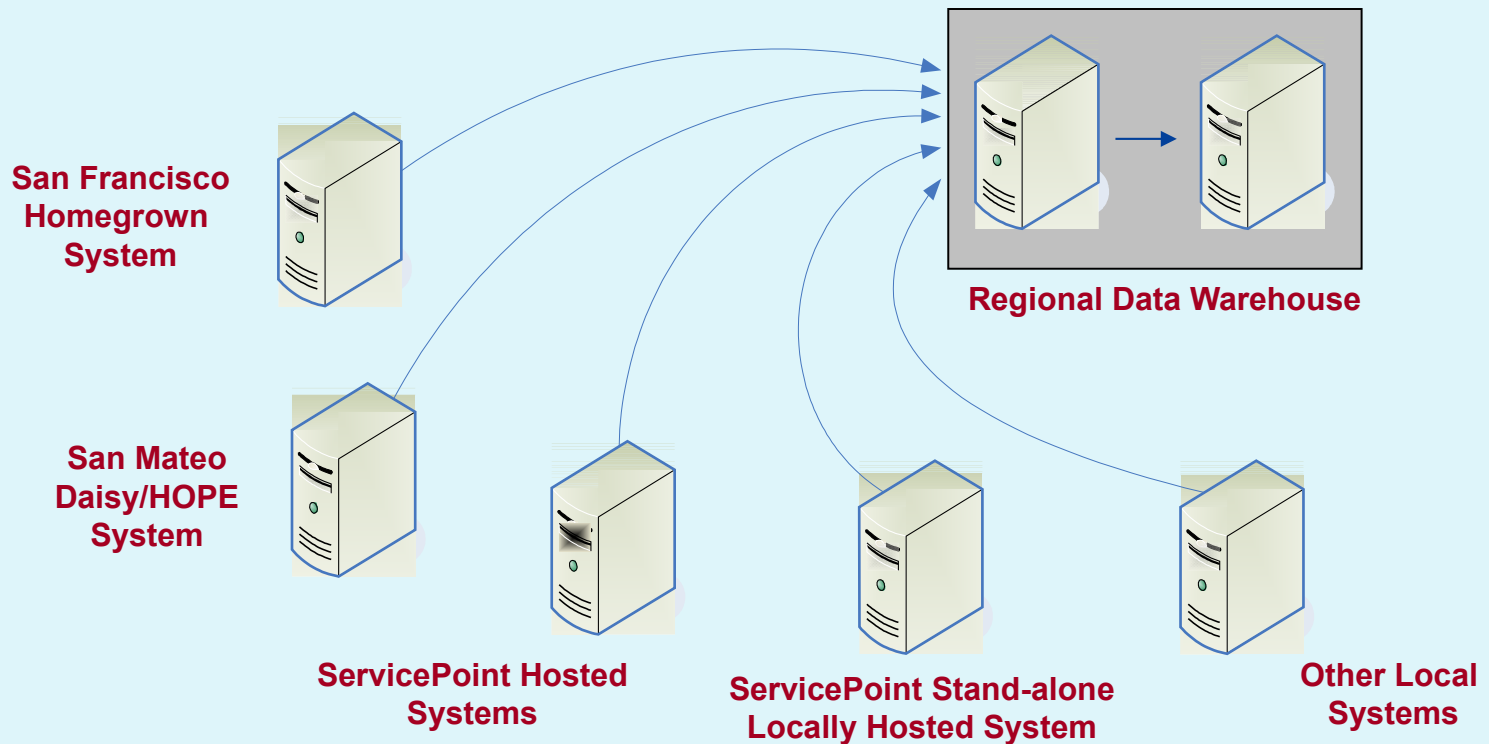
20

- **Bay Area Counties Homeless Information Collaborative**
 - **Mission:** To better enable policy makers, service agencies, and funders to understand and service the needs of the homeless within the community.
 - **Goals:**
 - ✦ Obtain 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
 - ✦ Leverage HMIS learning and expertise across multiple communities; increase success factors, reduce risk factors

REGIONAL DATA WAREHOUSE EXAMPLE: BAY AREA RHINO PROJECT

21

- RHINO overview



PLANNING AND PROCESS REQUIREMENTS

22

- Major planning challenges posed:
 - Planning may take longer than expected.
 - People aspects are often more challenging than the technology.
 - Start with the end in mind.

PLANNING AND PROCESS REQUIREMENTS

23

- Building a data warehouse starts with building a planning process:
 - Who should be at the table and why?
 - How often should the group meet?
 - What are the expectations for group members?
 - What is the work plan and timeline?
 - How will decisions get made?

PLANNING AND PROCESS REQUIREMENTS

24

- Who was at the table for BACHIC and SHADoW?

BACHIC

- Coordinators from 11 CoCs:
Alameda San Mateo
Contra Costa Santa Clara
Marin Santa Cruz
Monterey Solano
Napa Sonoma
San Francisco
- Lead Agency

SHADoW Leadership Board

- 6 state agency representatives
Housing Development Authority
Michigan Coalition Against
Homelessness
Human Services
Community Health
Corrections
IT
- Statewide HMIS staff and 3rd-party service contractors

PLANNING AND PROCESS REQUIREMENTS

25

- It is **critical to take the time needed to define core understandings and agreements.**
- Who is responsible for what?
- BACHIC core understandings and agreements:
 - Project Overview and Guiding Principles for RHINo
 - Policies and Procedures for Implementation and Operation of RHINo
- SHADoW core understanding and agreements:
 - Interagency Agreement
 - Participation Agreement
 - Data Use Agreement

PLANNING AND PROCESS REQUIREMENTS

26

- Building group consensus is critical!
- Three keys to success:
 - **Make a compelling case for the value of the regional or statewide data including a vision of the benefits from all the perspectives at the table.**
 - **Actively address participants' specific concerns and problems.**
 - **Allow sufficient time and dialogue to support relationship building and trust.**

PLANNING AND PROCESS REQUIREMENTS

27

- To summarize - A well coordinated planning process is critical to success. Process matters!
 - Funding for planning and staff
 - Define participants roles in the process
 - Identify barriers to success and strategies to overcome them
 - Identify key champions. Technical, programmatic and political
 - Work by consensus whenever possible. Support with democratic voting
 - Make program and policy decisions before technology choices
 - Document decisions and procedures
 - Consider project sustainability

PROGRAMMATIC REQUIREMENTS

28

- Think backwards
 - What is the purpose of the system?
 - What must the system be able to accomplish?
 - What sort of analysis and reporting is needed?
- What do we need and how much is it worth to us?

PROGRAMMATIC REQUIREMENTS

29

- Start with an overall Vision/Mission/Goals.
- The Vision/Mission and goals will drive all the other planning and program requirements.
 - ✦ Example: Because the RHINo vision included analyzing homeless “mobility pattern,” the last permanent zip code was included as a data element in RHINo as a means for tracking mobility.

PROGRAMMATIC REQUIREMENTS

30

- Identify specific data elements needed to address overall Vision.
- Ensure consistency of data (naming protocol for data elements).
- Ensure quality of data.
- Have a method for de-duplicating the data.
- Establish who “owns” the data and who has access to the database.
- Secure transmission of the data.

PROGRAMMATIC REQUIREMENTS

31

- Utilize off-the-shelf software or shareware to minimize costs
- Do not undersize hardware requirements – assess size of data
- Plan for scalability (future growth)
- Consider technical abilities/sophistication of those providing the data
- Develop protocols for up-dating and additional data sets
- Document **EVERYTHING!**

SYSTEM DESIGN CONSIDERATIONS

32

- Design will be driven by overall Vision, Mission, and Goals set early.
- What you want to know will dictate:
 - Data elements to be collected
 - Types of reports to be generated and schema set-up
 - Granularity and periodicity of data dumps
 - Equipment you will need
 - Security and privacy that needs to be put in place
 - Technical expertise needed at both ends (export and import)

SYSTEM DESIGN CONSIDERATIONS

33

Firewall



Upload Server



DW Server



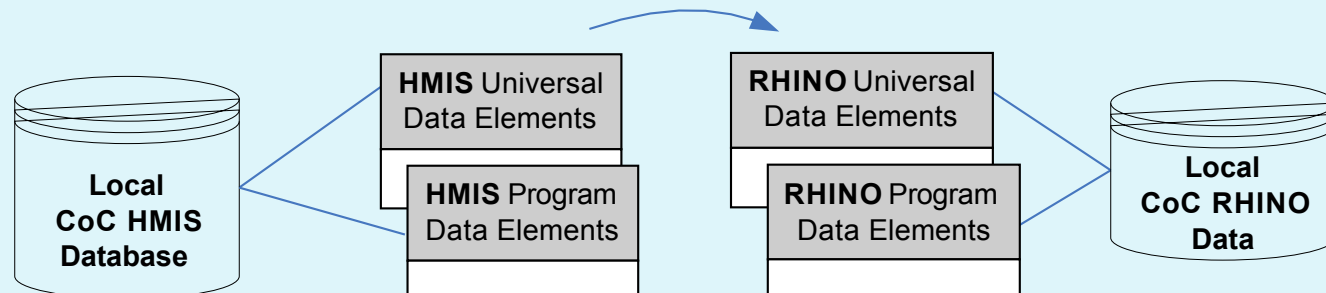
Backup Drive

- Minimize source system efforts
 - Especially ongoing duties/obligations
- Security, privacy
- Multiple diverse HMIS/mainstream systems
 - Different stages of implementations
 - Different data formats
- Reporting
 - Flexibility for future reporting choices
- Work flow
 - Procedures for resolving exceptions

INFORMATION PROCESS ISSUES

34

- How to get from “A” to “B” – export from disparate systems and import into the data warehouse. RHINO example:



- ETL (Extraction, ~~Transmission~~, Loading)
 - Consider technical sophistication of staff involved before deciding on ETL process
 - XML verses CSV
 - ✦ Extensible Markup Language (XML) verses Comma Separated Value (CSV)

INFORMATION PROCESS ISSUES

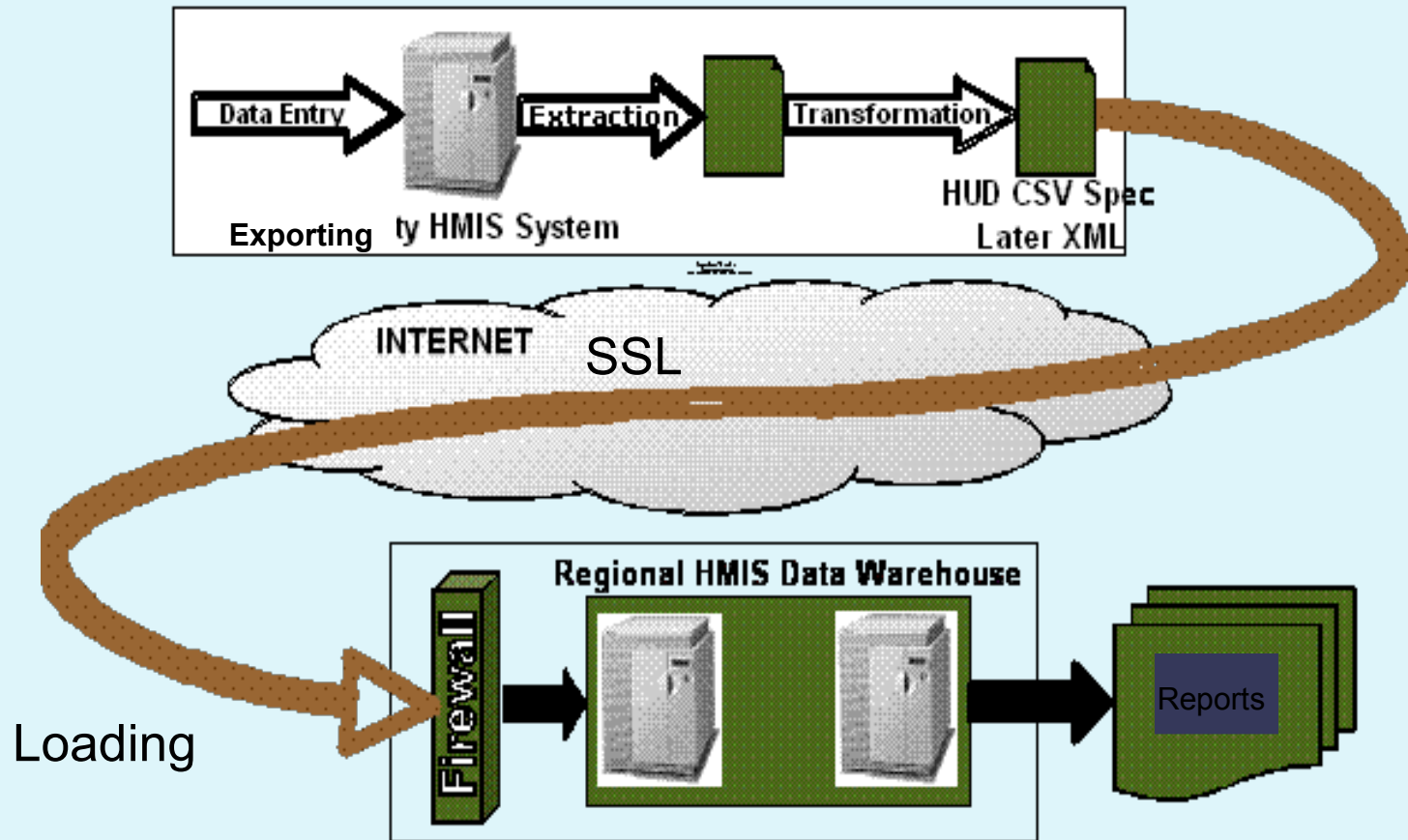
35

- Each exporting system formats its data into agreed standard structure (XML or CSV).
- Map data - create Mapping Tables:
 - Ensure information in all exporting systems and importing system mean the same thing. An “Apple” in one system is an “Apple” in all systems.
 - Protocols for mapping are in place, e.g., all systems use 8 digits for date of birth (03/07/1962).
- Ensure security of data in transmission:
 - Use SSL or SSL2 (Secure Socket Layer) encryption when transmitting the data from one system to another

INFORMATION PROCESS ISSUES

36

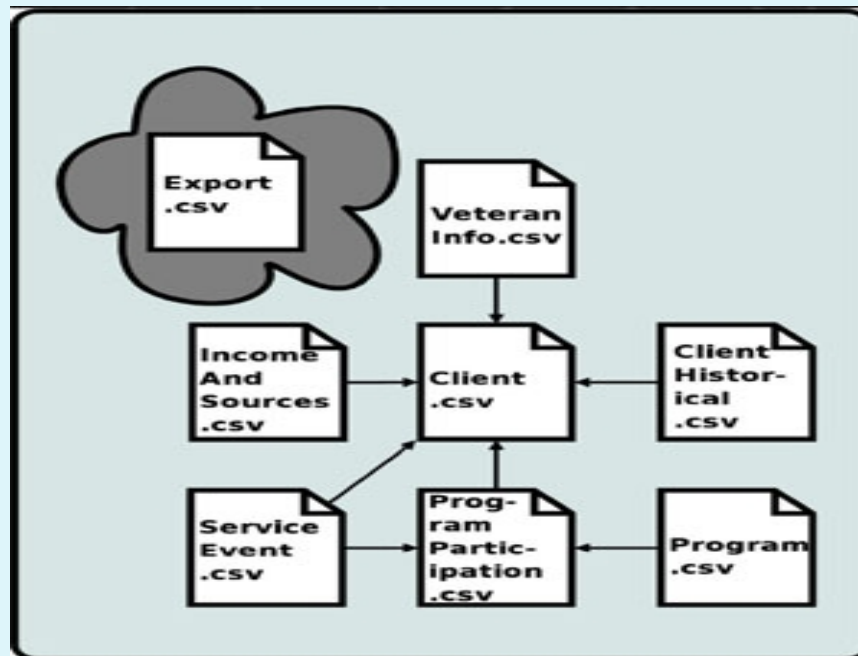
RHINo process example



INFORMATION PROCESS ISSUES

37

- Loading or importing into the warehouse:
 - Data should be re-organized to facilitate easy reporting on those things you want to know from Vision/Mission/Goals.
 - Most data warehouse structures use a “star schema.”



PRIVACY AND SECURITY ISSUES

38

- The protection of client information should be paramount.
- If personally protected information (PPI) such as name, SSN, date of birth are not being transmitted to the data warehouse then there has to be a mechanism for being able to de-duplicate the records.
- 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.

PRIVACY AND SECURITY ISSUES

39

- Security features should include:
 - Encryption during data transmission and storage
 - Firewalls
 - Anti-virus software
 - Spyware
 - 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

PRIVACY AND SECURITY ISSUES

40

- Using a Unique Identifier (UID):
 - Data security and data de-duplication are complementary rather than competing objectives.
 - 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 for use across exporting systems.
 - Uses a hash algorithm to encrypt UID.
 - Generate 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.

SIGNIFICANT CHALLENGES

41

- “People”
- Creating buy-in at all levels
- Funding and staffing
- Building consensus around a joint Vision
- Agreeing on data elements to collect
- Data quality
- Data ownership
- Management of multiple technical vendors
- Protecting privacy and confidentiality of client records
- Agreeing on format for reports
- Consensus: moving at the speed of the slowest partner

KEY SOLUTIONS AND LESSONS LEARNED

42

- People issues are more challenging than technical issues:
 - Build consensus at an early stage around a common vision – focus on what the system must be able to accomplish.
 - Technology serves policy, not the other way around.
- Building buy-in and consensus can take much longer than expected:
 - A project “champion” is an essential ingredient.
 - Augment the consensus process with democratic voting.
 - Imperative that overall Vision supports and does not compete with local interests.

KEY SOLUTIONS AND LESSONS LEARNED

43

- Basic understandings about participation need to be clear and documented.
 - Document EVERY decision.
- Independent outside facilitation can help move the process forward. No vested interests are perceived.
- **EVERYTHING** takes longer than planned.
 - Ensure that a structured planning process is in place – **PROCESS MATTERS.**

KEY SOLUTIONS AND LESSONS LEARNED

44

- Use the Vision to reach agreement on data elements to collect.
- Make data quality a responsibility of the exporting system.
 - Do not attempt to change data –it is not your data!
- Ensure there is a clear understanding on data ownership and what can be done with the data.

KEY SOLUTIONS AND LESSONS LEARNED

45

- Have clear protocols set around who has access to warehouse and data contained therein.
- Have clear protocols around type of reports that are generated and how they are “published.”
- Ensure that all partners are comfortable with security and confidentiality of the data.
- Achieving consensus takes time.
 - Make sure you take the time!

DATA WAREHOUSING: ADDITIONAL RESOURCES

46

- **Curricula:**
 - Data Warehouse Planning and Governance
 - Data Warehouse Program Requirements
 - Data Warehouse System Design and Technology Choices
- **Sample Documents:**
 - BACHIC Overview & Guiding Principles for RHINo
 - SHADoW Interagency Agreement
 - SHADoW Participation Agreement
 - SHADoW Data Use Agreement