

# REGIONAL & STATEWIDE DATA WAREHOUSING

## *Data Warehouse*

### *Program Requirements:*

*Regional & Statewide HMIS/Human Services Projects*

*An intermediate curriculum*



On the Horizon:  
Expanding the Uses of  
Human Services Data Systems

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.



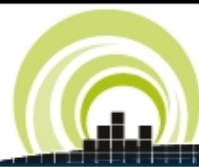
# LEARNING OBJECTIVES

- To introduce the audience to:
  - Establishing the vision and building a funding model.
  - Understanding some basics for staffing, and budget.
  - Demonstrating how privacy intersects with database design and output.
  - Providing an overview of key programmatic decisions regarding:
    - \* Key concepts for merging data;
    - \* Selecting the right information for inclusion from the right source;
    - \* Choices for moving and refreshing data; and
    - \* Planning for data use - access rules, publication policies, and reports.



# TRAINING OVERVIEW

- Data Integration
- Data Warehouse Background
- Visioning
- Adding Value
- Example of a Warehouse Cast
- Funding Models
- Privacy Issues
- Selecting Data Elements
- Access Rules
- Common Analytical Strategies



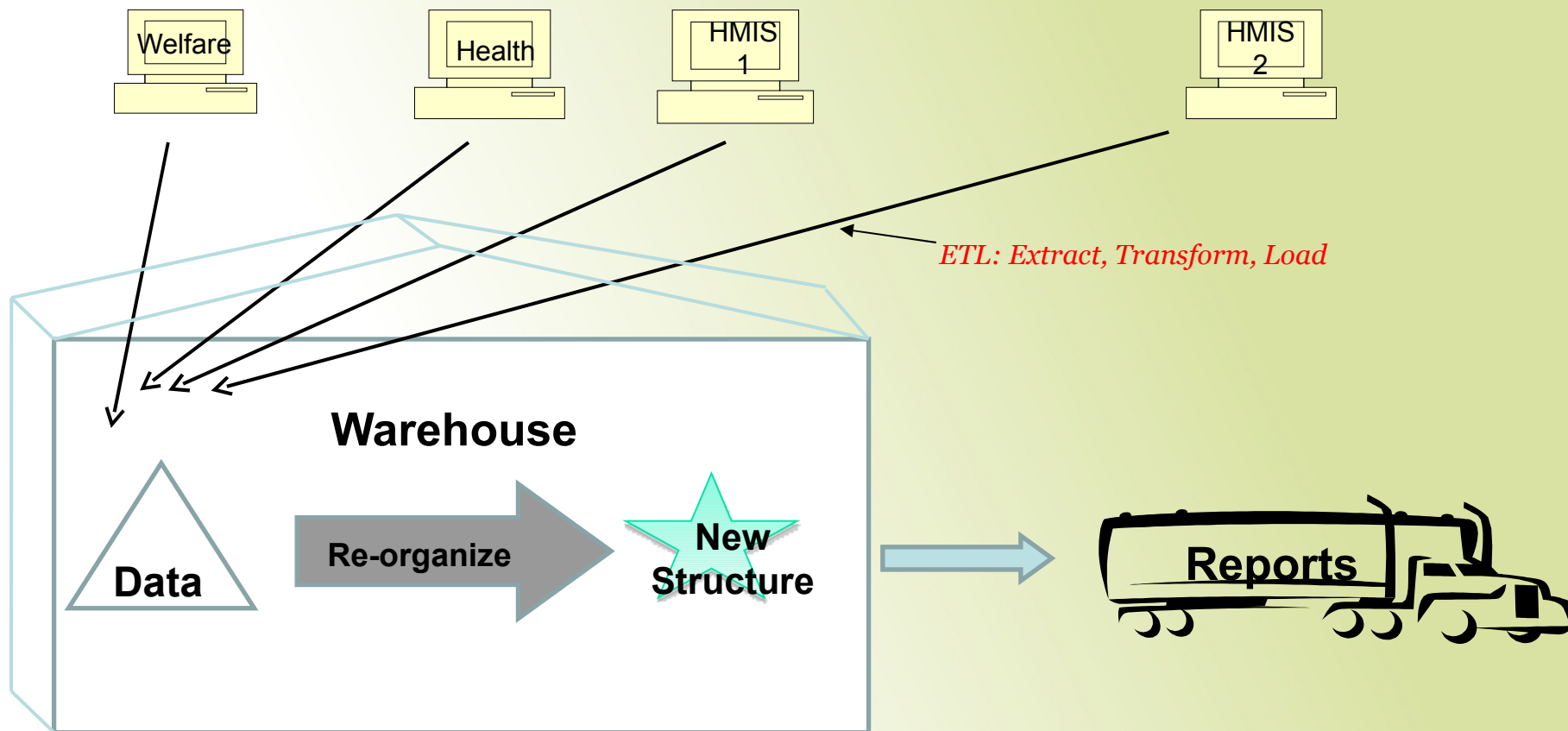
# DATA INTEGRATION

- 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** – Extensible Markup Language (XML) is a set of rules for encoding documents electronically for sharing a common case file for clients
    - Example – Michigan’s Muskegon project
  - Combining Systems** – merging data from 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?

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





# COMMON CHARACTERISTICS OF A DATA WAREHOUSE

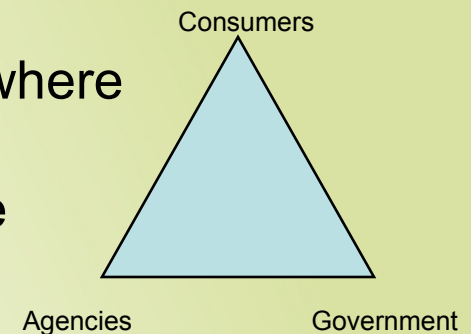
- The combination of data from two or more data sources into a repository to accomplish specifically defined analytical task(s).
- Vary from a very limited data set to detailed data depending on the defined task(s).
- The data is re-organized to support data mining (queries), reporting, benchmarking and analytics. An analytical tool such as Crystal Reports or Business Objects XI connects the user to the data.
- The warehouse also includes metadata (information about the data such as field type, length or source).



## THE VISION:

### Common Benefits Associated with Warehousing

- Warehouse partners will try to accomplish one or more of the following:
  - Using the warehouse as a tool to improve client services through improved cross jurisdictional / department coordination.
  - Evaluating the intersection of Systems of Care and/or geographies in order to answer specific research questions.
  - Reporting to funding organizations, planning, evaluation, or other basic business processes where data is located in multiple systems.
  - Avoiding burdensome double entry into multiple systems by exporting existing data.





# THE BALANCING ACT: From Vision / Benefit Analysis to Funding Model

- It is important to know why each partner came to the planning process.
- Prior to those initial meetings, project leaders should think through how the warehouse can benefit each partner. Frequently the benefits will be different for different partners.





## THE BALANCING ACT:

### From Vision / Benefit Analysis to Funding Model

- Warehouse planners may need to narrow the scope to create a cost/benefit balance. However, narrowing the scope too much may reduce the options available for funding the warehouse.
- Consider adding more value to your effort. With minor adjustments and appropriate organization, the warehouse can achieve multiple benefits, thereby creating more buy-in, supporting key analytical/reporting needs for partners, and increasing avenues for initial or ongoing funding.



# BENEFIT TO POTENTIAL FUNDING MODEL

## Warehouse Benefits

1. Improving services to clients through cross jurisdictional/ department coordination.
2. Answering key research question(s).
3. Reporting to funding organizations, evaluation, and planning activities.
4. Avoiding double entry, reducing staff time spent in record keeping.

## Potential Funding Sources

Private Foundations, As part of costs to a directs grant such as HPRP. Cost offset by savings.

State or Federal Research partnerships and Private Foundations.

Operating, planning, or evaluation budgets from a diversity of organizations from state agencies to large social service agencies.

Agency operating budgets. Costs offset by savings.

State/Federal/Private Foundation Grants may support initial design and other sources ongoing operations.



# ADDING VALUE TO THE WAREHOUSE: EXAMPLE 1

**Initial Goal:** Multiple HMIS jurisdictions combine data to study mobility patterns and create an unduplicated count. The data set includes demographics and entry/exit or service (transaction) dates and types.

**Transformation:** By adding a few more data elements such as “program type” and discharge status questions, additional tasks that might be accomplished.

1. Large multi-jurisdictional organizations need to generate data for funding organizations that reflect services across jurisdictions. This reduces the need for additional databases and related double entry.
2. Data warehouses frequently involve enough program and population diversity to allow the calculation of data-informed targets for funders without compromising individual agency privacy.
3. States use the system to generate required cross jurisdictional federal funding reports (ESG/IDIS/HPRP).



# ADDING VALUE TO THE WAREHOUSE: EXAMPLE 2

**Initial Goal:** Multiple CoC data is combined with mainstream data to study the cost of homelessness. The data set includes demographics, housing status, transaction dates and types and costs.

**Transformation:** The export tool will need to be updated to accommodate any new “test” questions (option 3 below).

1. Rapidly evaluate and report to the legislature the impact of program cuts or changes, by monitoring match rates in the private safety-net agencies participating in the HMIS.
2. Identify gaps in services related to specific poverty programs. Are WIC funds supporting homeless families? Do people receiving emergency homeless prevention subsequently become homeless?
3. Without modifying expensive state systems, the warehouse may be used to compile information on emerging issues quickly and efficiently sampling across jurisdictions. A sample of CoCs/agencies add new question to the HMIS for 3 months.



## AN EXAMPLE OF A WAREHOUSE CAST:

- The General: At least one dedicated staff to insure forward momentum, that coordination is maintained, and that deliverables are accomplished on time.
- The Board: Partner Leadership and other “experts” that negotiate the vision, make fundamental privacy and data use decisions. Plan for a series of meetings.
- Technical Support: Technical staff with significant allocation of hours during the warehouse design and build process and periodically as updates are needed to address changes in the source systems or scope.



# AN EXAMPLE OF A WAREHOUSE CAST (CON'T):

- The Planners/Data Users: Program/Evaluation staff during the design phase to define the specific data plan and report development. This group often evolves into the “users.”
- The Analyst: Be sure to include an Analyst to support report/query development, and also to teach those with access to use the reporting tools.



# COST/BUDGET CONSIDERATIONS

The design and operational costs will vary with the complexity and size of the project. Some important considerations:

- Ability to assign ownership to a key partner. Software Licensing and server costs may be reduced significantly by hosting within an existing IT operation.
- The ability to manage development time by very carefully completing the technical specifications. Limit course changes where value is uncertain. *Miscommunication and failure to manage can be very costly.*



## COST/BUDGET CONSIDERATIONS (CON'T)

- Cost will be highest during the initial design phase, but you must also prepare for hosting and updating.
- Ability to automate as much as possible. The more “hands-on” time required to construct each build, the higher the ongoing costs.
- Don't skimp on your Analyst time as the success of the warehouse will hinge on the utility of the “views” and the ease of mining the data, reports and related training.



## FINAL NOTE ON VISION

- Consider what is possible:
  - It may be possible to build on an existing warehouse leveraging previous investments by states, counties, cities, or agencies.
  - A successful warehouse is one where the benefits for all parties are clearly defined as part of the vision. Addressing one or more core processes for one or more partners will support a funding strategy.
  - Privacy questions will need to be addressed from the beginning and may limit the scope of the final product.

***The scope may change naturally as the details become known.***



# HOW PRIVACY INTERSECTS WITH DESIGN

It is not just about the technology!





# PRIVACY AND DE-IDENTIFIED DATA

## De-Identified Data Sets:

- HIPAA provides for disclosure of de-identified data for purposes of research and evaluation.
- De-identification involves not just the removal PPI (name, SS#, date of birth); but also additional information that could be used to identify a client from secondary sources such as Voting Roles. Some examples of risky data are:
  - Contact information such as addresses or phone numbers
  - General location such as city, county, or neighborhood may disclose a specific person when it is coupled with low probability characteristics.
  - Dates of services may also provide disclosing information. Whether a specific date is revealing depends on how rare the service is.
  - *The size of the dataset is important.*



# POTENTIAL USES OF DE-IDENTIFIED DATA

- Compare homeless characteristics across jurisdictions in the context of systems of care. For example, is it different for persons in families and for singles and does the service mix impact the patterns?
- Create multi-jurisdiction funding reports for large organizational partners.
- Identify the movement patterns of homeless persons between jurisdictions.
- Determine the success of prevention dollars by tracking re-admission to shelters across multiple jurisdictions.



# POTENTIAL USES OF DE-IDENTIFIED DATA

- Determine the impact of public policy changes on the “not for profit”/private system of care. For example,
  - Does a particular state cut increase the number of people entering shelters?
  - Does a change in poverty guidelines exclude persons who may have previously benefited from a service?
- Determine how local police enforcement practices affect homeless singles and families.
- Analyze the characteristics of prisoners released from jails that become homeless.

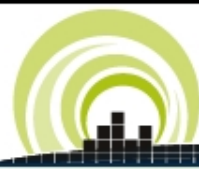
*And many, many others.*



# PRIVACY AND IDENTIFIED DATA

Identified Data Sets are those that include personal protected information (PPI) such as names, SS#, date of birth.

- The rules are more restrictive rules because the potential for harm is much higher. *See HUD Data Standards and HIPAA for guidance.*
- Many projects will involve an informed consent process which impacts the amount of historical information available.
- Research data sets will also require IRB (Institutional Review Board) approval to evaluate the risks to consumers.



# PRIVACY AND IDENTIFIED DATA (CON'T)

- Within the HMIS world, agencies frequently have concerns about releasing identified information even if the law provides for such releases under certain conditions.
- Warehouse designers should seek legal guidance as multiple laws often apply.

***A great deal of care needs to be taken in the use of identified data sets for numerous reasons.***



# POTENTIAL USES OF IDENTIFIED DATA

- PPI will be needed for Coordinating Care in order to:
  - Identify for the purposes of providing additional benefits, families on TANF who are not benefiting from WIC.
  - Identify families on Food Stamps who qualify for TANF.
  - Assure the HPRP funds are coordinated across neighboring jurisdictions and with other funding sources.
- PPI is also required to conduct research where:
  - Direct client input is required. For example, longitudinal research where client interviews are a critical step and the client is followed across time.
  - The product of the research will impact services provided to individual clients such as a medical study within the HOPWA (HIV/AIDS) population.

*May require informed consent from the client and careful review by an Institutional Review Board (IRB).*



# MERGING RECORDS AND PRIVACY

- Combining data sets from multiple source systems requires the system to match records. Which records are about the same person?
- Matching the records to accomplish the merge requires identifiers even in de-identified data sets.
- Planners will need to provide for the matching process within the privacy guidelines which usually involves a *“trusted” party* with access to the complete data set.





# SUGGESTIONS FOR ADDRESSING PRIVACY

- Privacy Notices must address planning and research in the “Uses of Data” section of the Notice.
- Participation and Data Use Agreements should include the details. Negotiate who will house the data - your “trusted” party.
- Do not store PPI for de-identified data sets.
- Define an IRB process in anticipation of research uses.
- It may be a good idea to include a process for some agencies to “opt out” if the integrity or size of the data set allows for less than 100% participation. Work is always easier with the willing and engaged partners may allow for enough scope to expand funding options.



## IDENTIFIED VS. DE-IDENTIFIED DATA

- Data Warehouse designers should always seriously consider whether their objectives can be accomplished with de-identified data.
- Some steps in privacy processes may not be about what is legal, but rather what will establish and maintain trust between partners.



# BEYOND PRIVACY— The Data Ownership Discussion

- Designers should be prepared to discuss the question:  
“Who owns the data?”
- While frequently included in contract/MOU language, it is important to note that determining ownership is not simple.
- A Lawyer may facilitate this discussion.  
*Ownership is not an absolute. It is really a question of who has what rights and who has what responsibilities. Ownership is really a basket of different attributes, rights, and responsibilities.*



# ADDITIONAL PROGRAMMATIC DECISIONS





# SELECTING THE DATA ELEMENTS

- The parameters of the data set will be determined by the vision/objectives of the project.
- Be prepared to adjust the scope of the project as you achieve a better understanding of the strengths, limitations, and privacy issues of the source systems.
- Administrative databases typically include a lot of bad data. Program staff will be necessary to identify which data elements are subject to data quality screens and are collected reliably. Additionally, be sure that you understand the specific definition of the field and if that definition is stable.
- Finally, understand the context or intent of each source system as it will impact how the data may be used.

***Making sure that you map “like” to “like.”***



# SELECTING THE DATA ELEMENTS - MATCHING

- More and More about Matching. What elements will you need to match?
  - Does the matching process include an probability estimate that the match is true (e.g. - 90% likelihood the match is accurate)?
  - Understand the collection culture surrounding identity information. Do staff collect the name, DOB, and SS# in a consistent manner? Are name and DOB taken from personnel identification documents? Are consumers likely to provide random information?
  - A greater number of elements included in the match may not yield a stronger match depending on the collection culture. When at least one source system has weak processes, creating an intermediate algorithm may be the best choice.





# SELECTING THE DATA ELEMENTS – DATA QUALITY

- Once identity is established, planners will need to select which fields are added from which source systems.
  - The information available to include is negotiated by leadership and defined in Data Use Agreements. A process to update that agreement should be included to allow the Warehouse to evolve.
  - What is the right source where data is collected on common fields? Once again knowledge of collection culture is important.
  - How do you handle congruency issues? For example, HMIS System 1 indicates that a client is chronically homeless and HMIS System 2 indicates that he/she part of a family. Is this a matching problem, is this a change in circumstances, or does this reflect poor data quality? What filters should you build into the data set to stabilize definitions for data mining?



# SELECTING THE DATA ELEMENTS – THE HISTORY

- Planners will need to decide how to handle information that changes over time. Does the intent of the project create a need for the history of each response or is the most current response adequate?
- Planners should also consider the history of the source systems. Some systems may have known periods when the data was unreliable. This issue may limit the scope of the warehouse project.





## THE REFRESH CYCLE: How Frequently should the Data Set Be Updated?

- Planners will need to determine the updating schedule for the source data sets. Divergent systems have different entry and cleaning cycles. For example:
  - Michigan's HMIS found that the information is generally not stable until two months after initial entry as providers complete their various cleaning cycles. Data extracts are therefore designed with at least a 2 month lag.
- The intent of the warehouse will also determine how frequently data must be refreshed. Identified data sets used for care coordination will need "near real time" refreshes. In contrast, planning and research data sets may only need periodic updates.



## ACCESS RULES

- Who has access to the warehouse? Some warehouses may have very simple access rules restricting use to a lead agency. Some may have complex rules to allow for multiple uses and users of the data.
- What training is required to use the warehouse? Common analytical tools such as SPSS, Crystal Reports, BO XI, and even Excel or Access.
- Designers will usually establish multiple “views” of the data to optimize reporting.





# ACCESS RULES

- Views of the data may be specialized for specific questions and even specific partners.
- The simpler the data set or the more limited the scope, the easier these rules will be to establish. However narrowing the scope may also make it harder to find funding support.
- Be prepared to deal with trust issues throughout the process.





## ESTABLISHING RULES FOR USING THE DATA

- Developed by an Advisory Board:
  - Data uses are usually defined early in the vision setting.
  - Rules become more specific as the scope and detail become clear.
  - Consider different rules for different uses.
    - \* Queries to support dynamic planning efforts - not published outside of the partnerships.
    - \* Queries in response to requests for information from the public, reporters, or other social planners.
    - \* Queries to support fund accounting.
    - \* Reports prepared specifically for publication.
    - \* Research where causal conclusions may be drawn.
  - Discuss what happens if a study reveals unanticipated or possibly damaging results. This will involve a detailed review and hopefully, “a commitment to go with the substantiated finding” (i.e. the truth).



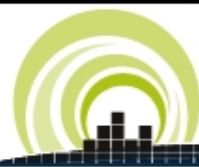
# REPORTING USING COMMON ANALYTICAL STRATEGIES

- Common analytic strategies assure that the data is interpreted in exactly the same way for all partners/jurisdictions, etc. All reports have multiple assumptions and calculation decisions. A warehouse supports the use of exactly the same report across multiple systems thus assuring a level playing field for decision making.
- Beyond views and a tool for querying the data, there are some core reports that should be planned. These are reports that are specific to the scope of the warehouse and where a common analytic strategy is important because the data will be used in a mission critical way. *Be sure that you understand the analytical plan.*



# DATA WAREHOUSING: ADDITIONAL RESOURCES

- 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



## FINAL MESSAGE

THE SUCCESS OF YOUR WAREHOUSE WILL NOT BE LIMITED BY THE TECHNOLOGY, BUT BY THE ABILITY OF THE PARTNERS TO TRUST AND ALIGN BEHIND A SHARED VISION.