Guidelines on the Use of Data Warehouses in Child Care and Protection Information Management and Analytics

September 2022

USAID FROM THE AMERICAN PEOPLE

DATA FOR impact

Guidelines on the Use of Data Warehouses in Child Care and Protection Information Management and Analytics

September 2022

Data for Impact
University of North Carolina at Chapel Hill
123 West Franklin Street, Suite 330
Chapel Hill, NC 27516 USA
Phone: 919-445-6949 | Fax: 919-445-9353
D4I@unc.edu
http://www.data4impactproject.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of the Data for Impact (D4I) associate award 7200AA18LA00008, which is implemented by the Carolina Population Center at the University of North Carolina at Chapel Hill, in partnership with Palladium International, LLC; ICF Macro, Inc.; John Snow, Inc.; and Tulane University. The views expressed in this publication do not necessarily reflect the views of USAID or the United States government.

MS-22-212 D4I

D4I is committed to local partner engagement and individual and institutional strengthening. Local authorship is important and we urge you to engage local partners in analysis and reporting.

USAID FROM THE AMERICAN PEOPLE

DATA FOR impact

Contents

Glossary

ad hoc reports	Creation of reports on an as-needed basis. Ad hoc reports are generally created for one-time use to find the answer to a specific question.
application log files	Documented activities of an application within the operation environment (i.e., server)
Business Intelligence (BI)	Using various sets of modern technologies to enable an organization lean towards data-driven decision making.
Central Database (DwH)	A central repository of integrated data from one or more sources. It will store current and historical data in one single place that can then be used for creating analytical reports for different audiences
cloud hosting	A data hosting approach hosted by a service provider in an offsite location, this could either be a public or private cloud
dashboards	A collection of charts, tables, and visual tools to represent a specific theme
data lake	A centralized repository that allows you to store all your structured and unstructured data
data mart	A simple form of a data warehouse that is focused on a single subject
data mining	Scouring data sources with an aim of generating relevant data
data models	A data model is a simplified diagram of a system and the data elements it contains, using text and symbols to represent the data and how it flows. Conceptual: a model that helps to identify the highest-level relationships between the different entities Logical: a model that describes the data as much detail as possible, without regard to how they will be physically implemented in the database Physical: a model that defines how data will be organized in a database
Extraction, Transformation, and Loading (ETL) Processes	Processes that extract, transform, and load data from multiple sources to a data warehouse or other unified data repository
encryption	Hiding true meaning of data using secret codes and keys
flat file	A collection of data that is stored in a two-dimensional database in which similar, yet discrete strings of information are stored as records in a table.
hybrid hosting	A combination of both cloud and on-premises hosting services.

metadata	Data that provides information about other data. Basic information might include purpose of the data, time and date of creation, creator or author of the data, location where the data was created, standards used, file size, source of the data, and process used to create the data.
No-SQL	Stands for “not only SQL” and refers to a non-structured database used for storing data in varied formats.
on-premises hosting	A data hosting approach where all the server hardware, firewalls, data, operating systems, and applications are physically kept in-house
query tool	An interface for looking directly at data and for constructing queries that can be used in quick reports
REST API	An interface that two computer systems use to exchange information securely over the internet. An application programming interface (API) defines the rules and Representational State Transfer (REST) is a software architecture that imposes conditions on how an API should work.
SQL	Scripting language for the Relational database management systems (RDBMS)

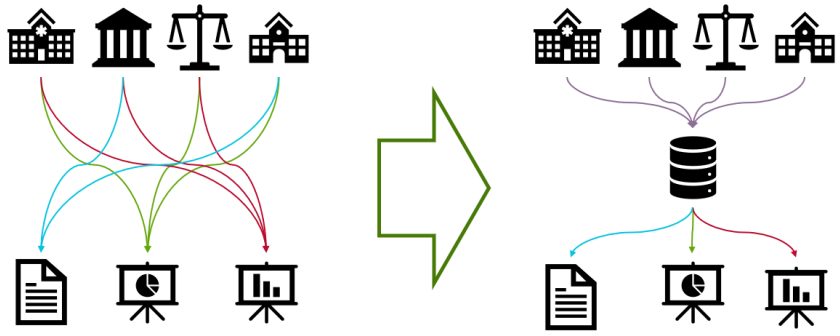
Introduction

Globally, countries are striving to reform child protection and care systems to ensure appropriate care for children without or at risk of losing parental care. Several instruments and global commitments inform these changes, such as the 1990 UN Convention on the Rights of the Child, the 2009 United Nations Guidelines for the Alternative Care of Children, the 2019 Resolution on the Rights of the Child, and the 2008 UN Convention on the Rights of Persons with Disabilities. **Consistent and reliable data are important and necessary to evaluate the effectiveness of these reforms both domestically and internationally, as well as to monitor the conditions of children in alternative care, understand risks and protection gaps that children face, and to inform funding, policy, and program decisions.**

Significant human and financial resources have been invested in the collection of data for decision making. In the child protection and care sector, data often come from a variety of sources and information systems owned and managed by a wide array of stakeholders from different sectors (social welfare, education, health, justice). Government and other stakeholders must typically consult several data sources and units within agencies to acquire the information they need. As a result, stakeholders consistently face challenges to accessing and using data to inform the design and implementation of child protection and care policies and programs.

Analytics solutions provide an automated way of mining data from a range of sources and visualizing it in tools such as dashboards to facilitate decision making and provide a single source for all stakeholders.

Figure 1. A data warehouse can streamline convoluted data reporting



There are a variety of possible analytical solutions, depending on the complexity of the needs and resources available. Each analytical solution will contain the following elements: data sources and metadata, a process to extract and transform that data, **ideally via a data warehouse**, and a tool for visualizing and reporting it out.

What is the purpose of this guide?

These guidelines were developed to support governments and other childcare and protection stakeholders in the conceptualization, development, and use of data warehouses to manage and analyze information from multiple sources for more efficient and informed decision making and

improved services and outcomes for children and families. It provides a technical overview of the options available for developing analytical solutions to integrate data using a data warehouse approach. For any level of complexity and budget, there are suitable solutions available; these guidelines aim to empower stakeholders with the key information needed for selecting the most appropriate one. Although many sections in the guidelines are intended for audiences with a monitoring and evaluation or information technology background, the document can also foster discussion among leadership and policy makers about data integration possibilities and goals for the future of childcare and protection information management and analytics.

How to use this guide

The factors determining how and when to implement a data warehouse will be unique to each country and use case. These guidelines serve as a starting point to understand the breadth of options available in terms of type of data warehouse, technical considerations, hosting options, and further to outline the types of additional resources and expertise needed to implement a solution. Each section includes an overview of the topic paired with a selection of external resources for deeper learning and instructions for implementation. This structure is intended to provide readers with the tools and knowledge to design the best process for their context, regardless of the starting point or final goals for the system.

Data Warehouses Overview

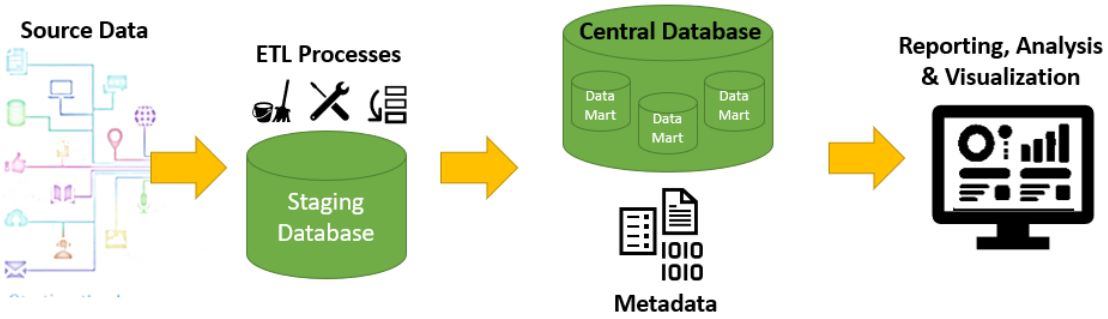
A data warehouse is a repository for large amounts of information pulled from multiple sources that can enable more comprehensive analysis for key actors and decision makers. In a childcare and protection context, a data warehouse might connect data from information systems dedicated to case management and child protection (like UNICEF's Primero), education, and health (like DHIS2). By storing data from these different sources in an organized format, a data warehouse can connect, manipulate, and transform the information it contains into a new "single source of truth" that analysts can query and use to develop one-off reports and continuously updated dashboards for different audiences.

Components of Data Warehouse

A typical data warehouse consists of four main components, though this can vary based on the levels of layers added in the design or the level of complexity desired. Variations on the primary components are described in more detail later in this guide.

- **Extraction, Transformation, and Load (ETL)** processes refer to the process of extracting data from its host sources, through a staging area where the data is transformed, translated, and reorganized to meet a specific data model before ending up into a structured data warehouse.
- The **Metadata** is the documentation of the data sources, usage, values, datatypes, and features of the datasets in the data warehouse. Metadata provides context to data and describes how to access data.
- The **Central Database (DwH)** may contain multiple databases of information organized inside of schemas, which you can think of as folders. When data is ingested, it is stored in various tables described by the schema. Query tools use the schema to determine which data tables to access and analyze.
- **Reporting, Analysis, and Visualization** tools connect to the central database, and are the window of the user into the data. They refer to reporting tools, structured and routine, ad hoc reporting structures, and connection to external websites for visualizations. They provide the ability to sort and organize data to answer specific questions and present it to end users.

Figure 2. Data Warehouse components



Determining Project Scope

As there are many potential use cases for a data warehouse system, it is critical to spend time defining the purpose of the system. Identify both the short-term benefits of the system, what its scope will be and how key stakeholders will use it, and the longer-term vision for how the system will affect services and ultimately, outcomes for children and families.

Potential use cases include:

- **Strategic Planning:** Inform the development or revision of policies and resource allocation decisions
- **Data Reporting:** Improve accountability with data used for federal, regional, and local reporting purposes
- **Performance Management**: Examine data to improve programs and the results achieved for children
- **Systems Strengthening Initiatives**: Assess the impact of policies and programs on students, workers, and education and workforce entities
- **Program Operations**: Access data to support day-to-day management and implementation of programs and increase their effectiveness

To select the most appropriate tools and avoid misalignment between what is designed and what was needed, there needs to be a common understanding of all the stakeholder needs the system will attempt to meet. Alongside this it is important to catalog the data sources that will be used and the resources available to implement them.

Stakeholder Mapping

Identify the stakeholders who will be supported by the data warehouse implementation. These are the end users whose needs the project aims to meet. This should include anyone who provides data for or will utilize the warehouse's outputs in their decision making. Be as specific as possible as the needs of each stakeholder may be different.

Many techniques exist for eliciting stakeholder needs. A few include: (1) *individual interviews*, which are helpful to understand specific view and needs, (2) *focus groups* that can be effective at getting the best understanding of how information flows between departments or teams, and (3) *use cases* can view through the whole system from a user's point of view, which can be useful in gathering functional requirements. Any combination of these may be used to develop a stakeholder list.

Table 1. Example stakeholder mapping chart

Stakeholder	Action/Decision	Data Source	Communication Channel
Key Questions	*What question does this stakeholder need to answer or what action are they trying to take?*	*Where does the information for this decision come from?* *Are there specific indicators they use?*	*How do they currently receive information and how would they like to?* *How often?*
Child Protection Committees			
Service Providers			
Regulators and Inspectors			
Researchers and Academics			
Administrators			
Government Officials			

Building a Data Ecosystem Map

A data ecosystem map can summarize major data activities in an ecosystem. Visualizing the scope, types of data being processed, stakeholders involved, data flows between different actors, and processes and platforms in use can help identify data gaps and possible duplication, support complementarity, and enable prioritization. This activity can be a helpful way to align technical and non-technical stakeholders and clearly show what will be included in the analytical solution, supported by a data warehouse.

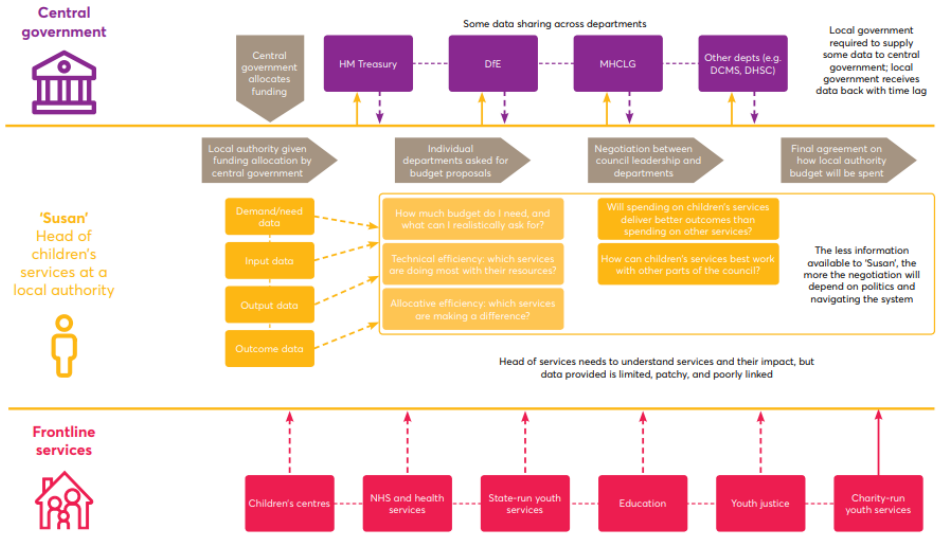
A conceptual data model should include all data sources related to the questions stakeholders are trying to answer, not just those that are currently digital. Its focus is understanding the business requirements of stakeholders, not how to best organize a database. Some key questions to ask at this stage include:

- What data assets are being managed in your context? (e.g., data on children in residential care facilities or kinship care, receiving grants or other social protection program support, case management, social service, and other allied sector workforce, etc.)
- What are the roles and organizations involved across the data management process for different activities? (e.g., caseworkers, local authorities, national authorities, judicial services, residential care centers, Information Management Officers, content experts, data providers, etc.)
- What are the relationships between these actors in the ecosystem? (e.g., regional office acts as facilitator, data sharing between implementing partners, etc.)

There is no standard template for a conceptual data model, and at this stage the most important function of the model is to lay out data needed to answer stakeholder questions and how they are related. At later stages of the project, additional details can be added to transform the conceptional data model into a logical data model. It can also be used to understand how non-digital sources interact with other data.

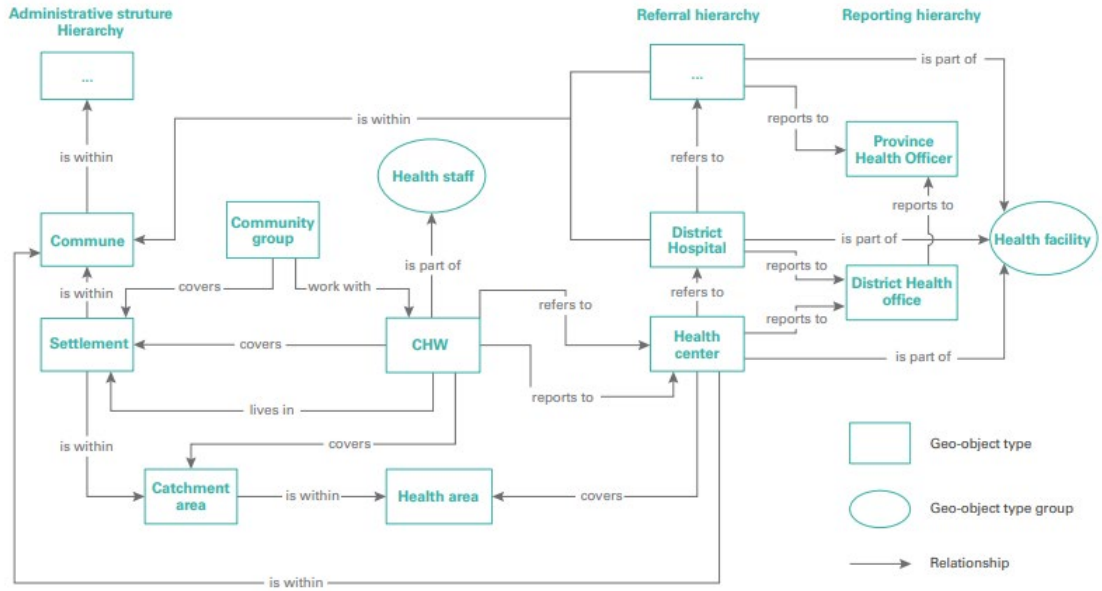
Below are two conceptual models representing data flows in child protection and community health work. Both serve as useful styles for gaining consensus on the data operating in the system while avoiding technical details.

Figure 3. UK Child Protection Data Flows[1]



[1] Figure originally used in 2020 Institute for Government Report Missing Numbers in Children's Services

Figure 4. Community Health Worker Registry[2]



The next step for each of the data sources identified in the conceptual model is to identify the details of the necessary data that are desired in the final analytical solution. Potential data sources may be administrative (collected as part of service delivery), management oriented (related to the financial or performance of services), or statistical and survey data which may be collected via survey or created in modelling exercises to represent the population.

Table 2. Example data source analysis

Data Source	Owner & Data Access	Data Updates	Format	Data Quality	Data Privacy
Key Questions:	*Who creates and owns this data?* *Who manages access? What process needs to be followed to get access?* *Are there any logistical or political constraints?*	*How often is data updated?* *How will we receive updates? Will updates be full replacement files or incremental*	*Excel file, SQL Database, etc.* *How much space does the data take?* *How complex is the organization*	*What, if any, standards does this data follow?* *Are there any important gaps in the data?*	*Has the data been anonymized to the greatest degree possible? Can data be aggregated?*
Data Source 1					
Data Source 2					
Data Source 3					

[2] Figure originally used in the 2021 UNICEF Report Implementation Support Guide: Development of a National Georeferenced Community Health Worker Master List hosted in a registry

Additional resources:

- Responsible Data for Children: Decision Provenance Mapping
- RD4C Data Ecosystem Mapping Tool

Data Source Suitability for Warehousing

Not all data sources are good candidates for an analytical solution supported by a data warehouse. The best data will be regularly available and updated, have a consistent format and be predominantly quantitative. The primary elements for assessing data quality are:

- Validity: Data should represent the intended topics
- Reliability: Data should have consistent collection and analysis methods
- Precision: Data should have sufficient detail to answer the questions
- Integrity: Data should have safeguards against errors or bias
- Timeliness: Data should be available frequently enough

Data quality and data standardization are important topics that fall outside the scope of this document, but every effort should be made to clearly articulate and document the data standards used by an analytical solution to ensure the smooth flow of data. Additionally, this guide does not cover data sharing agreements, which specify the conditions under which data can be shared between agencies.

Additional resources:

- NNIP's Resource Guide to Data Governance and Security
- USAID: How To Conduct A Data Quality Assessment
- Association of Public Data Users (APDU): Improving Administrative Data Quality for Research and Analysis
- CCEEPRA - Guidelines for Developing Data Sharing Agreements to Use State Administrative Data

Resource Mapping

The resources needed to develop the best solution possible will be both *human* (in the form of time available to dedicate to the project and technical skills that can be utilized), and *financial* (in the form of actual costs for software and hardware).

The profile of the team needed to implement a data warehouse depends on the technical solutions selected. A more basic or out-of-the-box solution will require less human resources than a customized or advanced one. Some general profiles will be required across implementations but may differ in the amount of time required or depth of skills. These include:

- *Project Sponsor*: A project sponsor should have the authority to make key decisions about the system and enact changes to allow it to move forward. This means the ability to commit budget and human resources to make the project a success, although they may not be involved day to day.
- *Project Manager*: A project manager is responsible for the overall project timeline and budget and may coordinate between stakeholders.

- *Data Architect*: The role of a data architect is to design and manage data systems. They develop policies for how data is stored and accessed, coordinate various data sources, and integrate innovative technologies into existing IT infrastructures.
- *Data Engineer:* The role of a data engineer is to collect and prepare data that will be used later by data analysts. They put into practice the policies developed by the data architect. They will be skilled in programming languages like Python, R and SQL.
- *Data Analyst*: The data analyst is responsible for taking the data in the data warehouse and analyzing it to create the dashboards and end-user products. They will be skilled in analytics, reporting and data visualization.
- *Software Developer:* Software developers design, program, build, deploy and maintain software, their expertise would be needed for more customized solutions.

When determining the financial resources needed and available, key questions to ask are

- What is the scope and audience of the system?
- What is the estimated budget range for the system?
- What budget will be available for ongoing support and maintenance?
- What is needed to prepare and approve a budget or resource allocation plan to fund the solution?

It is critical to remain realistic about resource constraints. It is better to identify a subset of the needs that can be accomplished and build on that success than to aim to accomplish all the needs and not be able to deliver any useful products.

Putting it all Together

The best way to avoid investing resources in a solution that does not meet the needs of users is to turn the mapping exercises above into a business requirements document. This document should clearly outline the

- *Project Objectives and Scope* – this should include the timeline, budget and deliverables that will be produced
- *Business Requirements* – these are the needs the project will meet, they should be described and ranked by priority, this will help clarify which requirements will be completed first or prioritized
- *Project Constraints* – this includes the risks to the project, resource to be used, dependencies to other work, deadlines, and budget

The stakeholder and data mapping process should be iterative and should be revisited as new stakeholders become involved or need change. Once stakeholders have access to information and begin utilizing it, they will identify new questions that can be answered with existing data and additional needs.

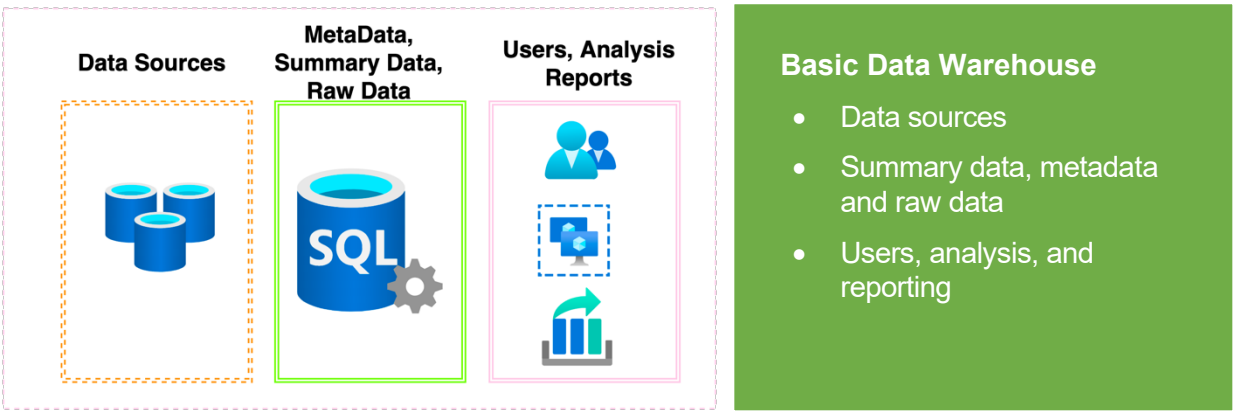
Selecting Analytical Solutions

The basic needs that every analytical solution will meet are; to provide access to data sources, to extract, load and transform data, and to see data in an easily digestible format. There are a variety of options available for both data warehousing and data analysis and visualization, and they can be combined in various ways to meet your needs. These options can also be reevaluated and substituted for more advanced tools as organizational needs grow and change.

Data Warehouse Options

Data Warehouse: Basic solution

Figure 5. Basic data warehouse



At the very least, a basic data warehouse architecture will have three components: (1) the data source(s), (2) summary of the data, and (3) users and reporting section.

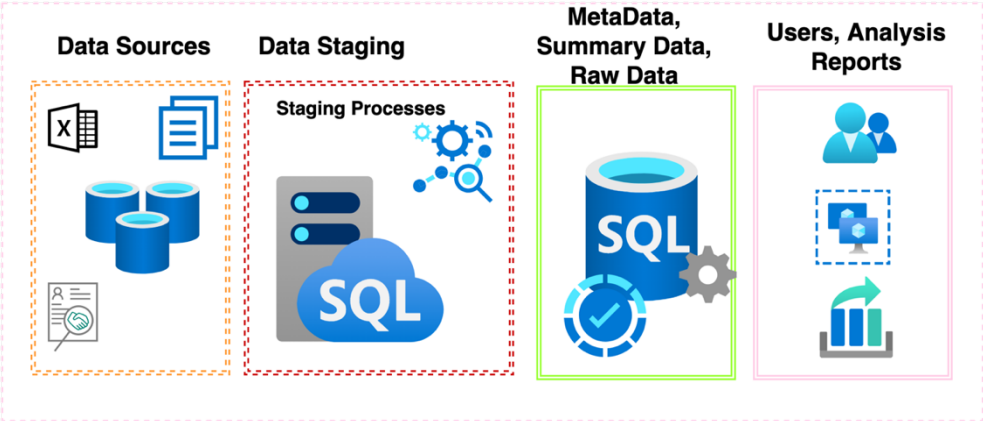
The Basic Data Warehouse option is appropriate when:

- There are budget constraints
- The scope of data focus is limited to one area of the organization
- There are few data sources, and they require very little transformation
- There are not many stakeholders who need to access this information

Data warehouse: Mid-level solution

The mid-level data warehouse is an improvement on the basic implementation. Here we introduce a staging area to ensure more processes (e.g., transformation or translation, calculations, etc.) can be run on the data before pushing to a summary database. These could be translation, transformation, pivoting or even data cleaning and data quality services.

Figure 6. Mid-level solution

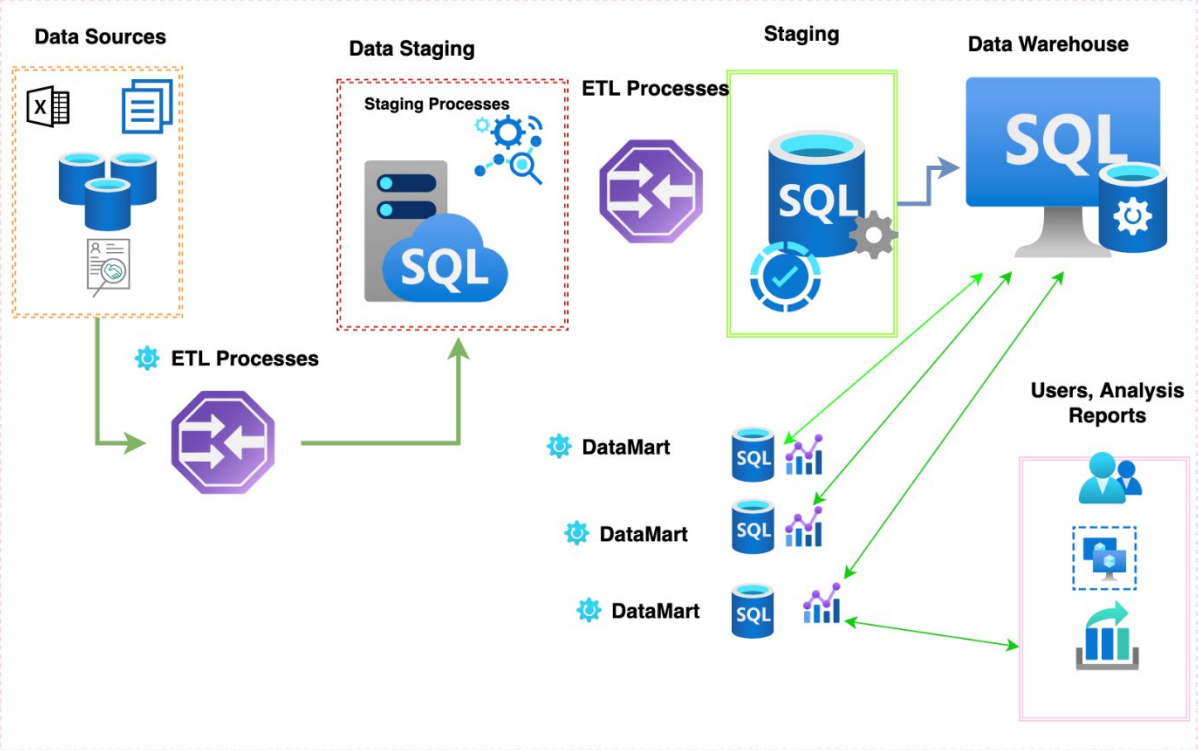


The mid-level solution is appropriate when:

- There is more budget and resources available
- Data formats are more complicated and require processing to be analyzed, or there are large amounts of data that would slow a simple system
- Scope is focused on one area of business for the organization (specialized approach)
- Minimal specialized services are required in the staging area
- Improve security, data compliance, storage capacity, integration compatibility, and shareability of data

Data Warehouse: Advanced solution

Figure 7. Advanced solution



An **advanced solution** will include additional features like data lakes, data marts, REST APIs for data exchange or an external app. It combines several functions or departments in a single unified database. Data marts can take a large data warehouse and break it into smaller and more tailored pieces for different audiences.

The advanced solution is appropriate where:

- Many users need various access information and there are distinct categories of data that they will each use
- There are large amounts of data, and some data may be unstructured
- There is sufficient budget to and skills to develop the needed infrastructure and plan for a more complicated implementation

Table 3. Warehouse summary

Basic data warehouse	Mid-level data warehouse	Advanced warehouse
• Specific to a smaller unit of business line • Usually smaller than the Enterprise Data Warehouse (EDwH) and Operational Data Warehouse (ODwH) • Offers subject oriented dataset	• Central database • Focuses on operational reporting, controls and decision making • A data source for the enterprise Data warehouse (EDwH) • Refreshed in real-time	• Spans across the entire enterprise • Collection of departmental databases • Presents a unified approach in organization and classification of data

Considerations for Selecting a Data Warehouse Model

	Basic	Mid-Level	Advanced
Technical Skills Needed	Minimal technical skills required. Data and spreadsheet knowledge is most appropriate for this.	Combined data and spreadsheet skills required. Experience handling database is highly desirous for this to be achieved.	Highly skilled personnel required. from visualization tools expert, data pipeline setup (ETL) experts, database expert and a software engineering skill will be required here.
Functionality	Minimal functionality required. Most utilizes existing functionalities.	Employs moderate functionality, might have minimal specialized functionalities to support some high-level demands.	Employs advanced and complex functionalities. Sometime built from scratch with the use of the specialized experts in the team.
User Access Management	Normal Laptop access and spreadsheet setup is sufficient.	A data staging area will improve security, data compliance, storage capacity, integration compatibility, and shareability of data	This solution is best when there are many users who need to access information and there are distinct categories of data that they will each use
Timeframe to Set Up	Can be developed rapidly (Takes 1-2 months to production)	Based on complexity of needs, requires 1–4 months to production	This option will require significant planning and will take the longest to complete. Requires 1-12 months to production.
Performance Considerations	This option is best suited for a limited number of data sources and a limited number of users. Too many users or too complex of data sources will cause this system to load slowly	A staging area for data will speed up the load time for preparing and analyzing data This option is best if focus on one type of data, fewer sources (more than one source but relating to one area of business)	Adding data marts will further speed up the load times for preparing and analyzing data .
Infrastructure Requirements	Demands basic infrastructure. This refers to a minimalistic investment in technology. Could easily be setup by a single workstation in organization and does not need external access.	Employs a mix of both basic and moderate infrastructure especially where distributed approach is embraced. Moderate expenditure in technology and data mining utilities, might have a need for external access.	High end infrastructure required to support high performance, access, load balancing and other specialized roles that might be in place. Complex setup and increased investment in technology from mining, ETL, access and third-party apps, might include high end servers.
Cost	This option has fewest minimal costs to setup	Moderate to high in terms of cost (based on tweaks)	Costly, though this might vary based on the choice of tools deployed i.e., high end servers, data mining utilities, third party apps licenses etc.

Data Analysis, Visualization and Presentation Options

There are many data analysis, visualization, and presentation tools available that suit analytical different needs and resource availabilities. These tools will connect to the data warehouse option selected and will serve as the interface with which users interact with that data. Connecting the data warehouse to visualization tools is done through a standard connector with a licensed BI tool, which can connect to either servers or files or with a REST API for more advanced deployments.

The key component of dashboards is who will be using them. A data warehouse allows the same data to be easily presented in different ways for different audiences and an efficient data warehouse will allow users to use a single interface with filters to get the information that they need to make decisions. For each information product you create, keep in mind the following key questions:

- Who is the audience for this dashboard?
- What action do they need to take with this information?
- How much time will they spend with this data and how often will they need it?

Each stakeholder may have different needs for the types of data to be shown, the granularity of the data, or the ability to drill down to investigate more. For example:

- A **case worker** might want to see data at a high level of granularity to make individual-level decisions, e.g., where to focus home visits that week.
- A **supervisor or administrator** in **subnational government** or organization providing social services may be interested in a more tactical dashboard, where they would look at data with a lower level of granularity, looking at trends in their geographic area over the past quarter or several months, e.g., proportion of children living outside of family care, receiving cash transfers, out of school, to better allocate resources or staff.
- A **department head, director, or policymaker** in **national government** will likely want a more strategic dashboard with aggregated data over a time series, e.g., to inform strategies and other planning and budgeting processes.
- A **donor** might be interested in a high-level view of the data to gauge the performance against the agreed-upon indicators and targets.

For each dashboard to be created, design a prototype to understand what data should be displayed and how the users prefer to see it. Outlining these needs will guide the decision-making process for which data visualization tools to use.

Data Visualization Options

An **out-of-the-box solution** would be to deploy a well-known business intelligence platform on-premises or in the cloud. Analytics would be conducted within the platform, and it would be individually licensed or embedded within an existing web application. This option is faster and easier to start up but will have ongoing license costs that may be prohibitive.

A **semi-custom solution** would be to deploy an open-source BI tool within a local hosted and managed environment. This option would mean that there were no recurring licenses to maintain, but would require more technical skill and offer more limited support.

A **fully customized solution** would be to develop a custom application using open-source or low-cost charting libraries. This option would mean that there were no licenses to maintain, but would require significantly more advanced technical skills on an ongoing basis for portal maintenance and integration.

Considerations for Selecting a Data Analysis, Visualization and Presentation Tool

	Out of the Box: Licensed BI Tools	Semi-Custom: Open-Source BI Tools	Fully Customized: Custom Solutions
Technical Skills Needed	Out-of-the-box set-up is easiest of assessed options Longer-term maintenance of the platform performed by the platform provider and is low burden for implementer User support is robust and supported by the platform provider.	Set-up is more involved than a licensed BI solution and requires installation of individual components onto on-premises server Maintenance costs would be higher because these options are newer and therefore have less documentation, training materials and support available	Set-up is like open-source BI tools and is more involved than a licensed BI solution and requires installation of software onto on-premises server Longer-term maintenance will require technically skilled staff with software development background
Functionality	Extensive functionality ranging from analysis through visualization Marketplace of third party-developed visualizations, open-source tools, and custom connectors to platforms like DHIS2 available out of the box	The suite of functionality isn't as wide as licensed BI tools Implementers can build components, but this requires technical skills, and there is no centralized marketplace with out-of-the-box third party components	Out-of-the box features for custom visualization software are limited as compared to BI platforms, and functionality would need to be extended through code
User Access Management	User access is available and built into the BI tool Ability to control users' access to specific subsets of data	Generally, more freedom because implementer has the power to structure things based on the architecture being used but requires a developer to define and implement.	Generally, more freedom because implementer has the power to structure things based on the architecture being used but requires a developer to define and implement.
Timeframe to Set Up	Fastest option for set-up (including installation and user creation) (~2 weeks)	Similar timeframe for set-up compared to licensed BI solutions User configuration will take longer because it is built by the implementer and not as an out-of-the-box feature in the open-source solution (~3 weeks)	Process will take several months, which will involve a process of designing, building, testing, and deploying the solution, and will required an iterative process to collect feedback to update and optimize the interface for accessing visualizations
Performance Considerations	Performance limits are based on server sizing and optimization rather than platform limitations Concurrent users would need to be monitored to determine performance constraints Licensed platforms use highly optimized data models, provide caching, and have properly sized servers to handle data processing	Compared to licensed BI tools, open-source options are typically newer and still face some bugs and issues that are unresolved The implementer will need to employ a software developer to optimize processing (whereas most licensed BI tools are built optimized to manage large amounts of data)	These custom tools would provide little to no optimization, and performance improvements would be driven by the lower-level design and be the responsibility of the implementer
Cost	Content creators and viewers receive a license directly from the BI platform provider: Long-term cost encompasses HR oversight in administrative user to maintain users and role definitions Less dependency on technical developer skills and technical oversight	Cost related to the technically skilled workforce required for set-up (does not use a per-user cost model) Long-term maintenance requires software development skills, and is higher cost over time compared to licensed BI tools	Low cost for Visualization license Will require cost of technically skilled staff to complete set-up and code analytics and visualizations Future maintenance encompasses yearly fee for license (low) and time/LOE from technically skilled staff with software development background (high)

Data Hosting

All the data in the warehouse and used in visualizations must be stored in a way that is secure and able to be accessed by the tools selected. This can be either on-premises, cloud or as a hybrid of both based on the implementation demands, replication or redundancy needs of an organization.

Key questions to consider when considering data storage options:

- Data security: Are local or cloud options up-to-date on current security regulations and standards? How will a data breach be handled? Are there any data types identified to be used that have different needs or standards?
- Bandwidth: Are there limitations related to uploading or retrieving data?
- Existing Systems: Any existing technologies in use that can be leveraged for the solution?

On-Premises Hosting

If hosting on-premises, all the server hardware, firewalls and software, operating systems, and applications are physically kept in-house — usually at the company's physical office. The organization hosting that physical equipment is then responsible for overseeing the server's wellbeing, security, and performance.

Cloud Hosting

If utilizing cloud hosting, the virtualized server is hosted by a service provider in an offsite location, this could either be a public or private cloud. This entails a third-party provider who physically hosts the data off-premises at a secure data center. The cloud hosting provider is responsible for all hardware costs, and server and firewall maintenance.

Hybrid Hosting

This mode is where the organization utilizes both the cloud and the on-premises for different types of data or parts of the data warehouse. Some applications could be running on the clous(offsite) and the security critical ones could be host on premise (onsite).

	On-Premises	Cloud	Hybrid
Pros	There is a lower total cost of ownership with no ongoing subscription costs Offers complete control over the configuration and processing any upgrades and changes Data that is locally stored can still be accessed in the event of network problems Sensitive data can be held locally and not transmitted outside of the organization	Maintenance of the software and hardware are covered by cloud service provider Data centers are likely to have data security teams more advanced than are practical for organizations Flexible to pay only for what is needed and scale as necessary	Easily balance the costs between cloud and on-premises. Allows for control where required as the choice of cloud or premise is reserved for the client to make. Perfect for implementing redundancy and replication setup. Sensitive Data can be held locally, and non-sensitive data can reside in the cloud

	On-Premises	Cloud	Hybrid
Cons	Responsibility for all maintenance and security, data backups, storage, and disaster recovery are beyond what most IT teams are equipped to support High initial cost with the purchase of hardware Longer implementation timeline requiring more logistics and planning	Access to data stored on cloud servers requires reliable internet access There may be fewer configuration options The upfront cost is lower, but overall long-term costs are likely higher with ongoing subscription costs	On -premises section of the hybrid configuration might present high costs of setup. On-Premises infrastructural challenges might impact the entire setup if not well maintained.

Final Recommendations

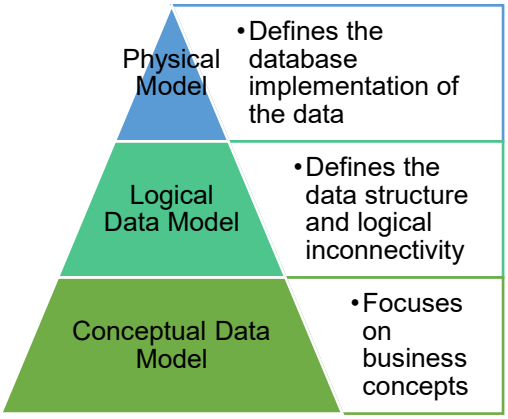
Each of the data warehouse options can be combined with any of the data visualization options and need not be considered final decisions. The most basic solutions can be iterated upon and replaced with more advanced ones as needs change.

With any version of implementation, it is critical to manage expectations, whether they be about a more basic system that many not meet all the stakeholder's needs, or a more advanced system that may take longer to implement than stakeholders expect. Maintaining close communication throughout the process will prevent misaligned expectations.

Implementing A Data Warehouse

The implementation process for the tools selected will vary significantly depending on the toolbox selected, but generally will entail the following components.

Figure 8. Data model types



Data Mapping

At this stage, the conceptual data model and during the stakeholder mapping can be elaborated into a logical and then a physical data model. This will contain details on the table names, column names and column data types that will be used in the data warehouse.

This process will also define which data sources will need to undergo a transformation to be used. Data to be moved must be defined, including the tables, the fields within each table, and the format that data should have after they are moved. For data integrations, the frequency of data transfer will also be defined.

This stage is another opportunity to consider the privacy and security implications of the data being used - for each data source you should determine if each of the data variables are justified by the scope of the project, and only collect what is absolutely needed. If data has been anonymized, consider whether through using multiple data sets, could the merging of these datasets be used to re-identify individuals.

Data Mining

This is also referred to as the data exploration stage. Three important activities take place here: (1) extraction, (2) transformation, and (3) load. Data sources are mapped out and mostly automated utilities (automated functions/routines) extract the database on a given criteria, load, and finally push to the staging database.

Data Processing

A staging database is a transition location for data. Here, further data transformation and translation routines can be conducted. These might include data transformation routines to ensure that data is in the right format or translation services to ensure data is in the correct language. These processes can frequently be automated routes that ensure the process is standardized.

Production

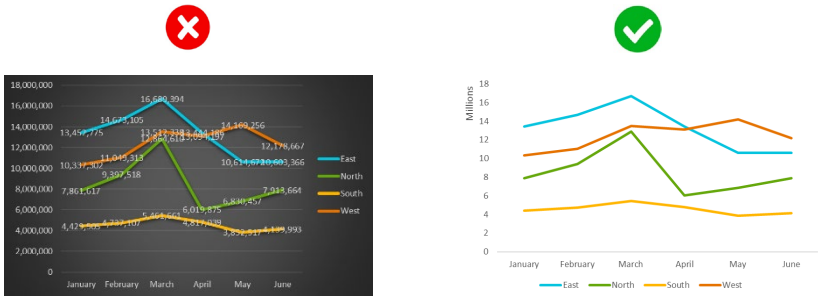
To ensure proper coordination of the changes and updates, the production pipeline needs to be set up to ensure that all updates are updated from the last mile approved changes to the production servers. Best practice is to have a test server (test server is a playground/sandbox for testing finished functions or items that are not yet loaded to the production server) updated with the latest changes, tested, and approved before pushing changes to the production server.

Dashboard Development

Designing effective dashboards is critical to achieving the purpose of the data warehouse. A data product can display the required information, but not be useful to its end-users because of design choices. An effective dashboard will simplify complex information and tell a clear story.

In dashboards and charts, you can improve the design is by reducing the amount of visual noise. The American statistician, Edward Tufte called this the *data-ink ratio*. The term comes from a pre-digital age when data visualizations were printed. Data ink is the ink that communicates data and non-data ink does not. Well-designed dashboards aim to use as little non-data ink as possible—in other words, they have a good data-ink ratio.

Figure 9. The data-ink ratio



Additional resources:

[ESRI - Designing Effective Dashboards]

[DataPine - Dashboard Design Principles and Best Practices]

Deployment

Deployment is typically the last phase of building a data warehouse. Assuming that everything is counter-checked, tested, validated for appropriate functionality, and passed Quality Assurance (QA), we can focus our attention on the following. Where possible, these validation processes should be executed on the test server before going live:

- Confirming that the infrastructure is in place and components are tested and ready to go
- The requirements checklist/change request document is validated for completion and sign-off
- Confirmation of the users, roles and access rights is in place
- The architecture validity is verified to be sound
- ETL Jobs loaded, tested, and verified to be optimal in performance
- Database is well defined with paging allocation articulately configured within the Relational Database Management System (RDBMS)
- Training users

Support, Maintenance and Monitoring

Once your analytical solution is in place, it is critical to offer support, maintenance, and monitoring to ensure that it remains operational and useful to stakeholders. Continuous iteration where necessary and change management becomes part of the maintenance and support process. What this means for your organization is something that you can define, but the following elements can guide your decisions:

- Training for new inexperienced users and ongoing consultation with existing users to identify new needs
- Regularly revisiting the stakeholder list and ensuring that new or changing needs are being met
- Regularly revisiting the data sources list to ensure that the best data sources are being utilized
- Define service level agreements with consultants

Although outside the scope of the data warehouse, an important project to consider, if they do not already exist within your organization, is creating Data Policies to outline the vision for responsible data retention, maintenance, and destruction. Depending on where you are, different laws may apply, and the aim should always be to exceed what is legally required. Data for children presents unique ethical and security challenges. The Responsible Data for Children Principles and Practices are critical to consider at every stage of the project to ensure that we are being appropriate stewards of children's data.

Additional resources:

- Data Protection Laws of the World
- Nethope Data Governance Toolkit
- Responsible Data for Children: Opportunity and Risk Diagnostic Tool

- UNICEF Good Practice Principles on Information Handling and Management in Child Protection Information Management Systems
- RD4C- 22 Questions to Assess Responsible Data for Children

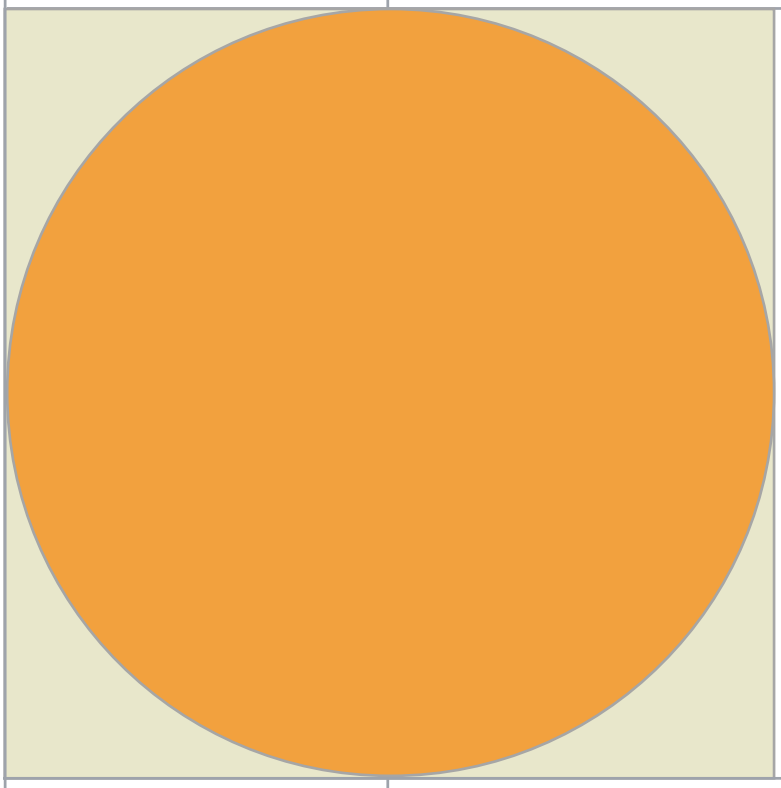
Data Security and Responsible Data Use

Since your data warehouse will combine potentially sensitive information from various sources, it is important to understand the security implications and develop a system that protects all the users' and clients' information. This includes minimizing data to be stored, protecting it in transfer through encryption, protecting the data warehouse itself, and ensuring that only the right users are accessing data through clear user groups and login credentials. The field of IT security is complex and best considered by a cybersecurity expert, but some key elements that should be examined are:

- Assess risk of data flows from collection to storage and distribution to identify points of vulnerability
- Examine the policies of any vendors being used and understand how they are preventing unauthorized access to data
- Develop or critique procedures for dealing with security breaches
- Establish a process to identify and recover lost, corrupted, or tampered data
- Establish oversight mechanisms and ensure that staff understand secure management, sharing, and transmission of data

Additional resources:

USAID Using Data Responsibly
Data School - Data Warehouse Security
Data for Children Collaborative: Ethical Assessment
Responsible Data for Children: Opportunity Risk Diagnostic Tool



Data for Impact
University of North Carolina at Chapel Hill
123 West Franklin Street, Suite 330
Chapel Hill, NC 27516 USA
Phone: 919-445-9350 | Fax: 919-445-9353
D4I@unc.edu
http://www.data4impactproject.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of the Data for Impact (D4I) associate award 7200AA18LA00008, which is implemented by the Carolina Population Center at the University of North Carolina at Chapel Hill, in partnership with Palladium International, LLC; ICF Macro, Inc.; John Snow, Inc.; and Tulane University. The views expressed in this publication do not necessarily reflect the views of USAID or the United States government.
MS-22-212 D4I

USAID FROM THE AMERICAN PEOPLE

DATA FOR impact