Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

Public Health Surveillance and Data

Improving Public Health Data Pipelines

CDC and partners are helping health departments make informed decisions by reducing the level of manual effort and time needed to access and use quality, analysis-ready data from multiple data sources.

Data pipeline pilot project

CDC and the United States Digital Service (USDS) co-led a pilot project with the Virginia Department of Health January through September of 2022. The project led to the creation of a prototype data processing pipeline that validates, ingests, and links data across multiple data streams so it can be used for timely public health action.

Problems with public health data pipelines

Data coming into state public health agencies can be messy. It arrives in multiple file formats, lacks standardization, and is often incomplete.

Processing this data so it can be linked across person, place, and time and used for action, like case investigation or emergency response, involves a lot of time and effort from health department staff. It can require coordinating many systems and often results in unnecessary duplication of work, delays, and differences in the data.

Detangling the data

> This work is part of CDC's broader efforts to help state, territorial, local, or tribal (STLT) health departments reduce the significant manual effort needed to access clean, analysis-ready data for public health action across multiple data sources and use cases.

The project team explored new approaches to storing, processing, and linking different incoming data streams to produce high quality, analysis-ready data and insights. Then, based on what they learned, they developed a prototype of a modern data processing pipeline for the Virginia Department of Health. They designed the prototype pipeline to help Virginia use lab, case, and vaccine data that they already receive so they can answer urgent COVID-19 public health questions with less manual effort.

The prototype

- Saves time and manual effort
- Increases data processing speed
- Creates a single source of truth for incoming data
- Removes the need for duplicative processes

Reusable solutions

Though the pilot focused on Virginia's needs, the project team used lessons learned to create reusable solutions that other state, territorial, local, or tribal (STLT) partners can use to solve similar public health data challenges. This approach follows CDC's blueprint for making public health data work better, also known as the North Star Architecture.

Data processing prototype

The pilot project team developed a data processing prototype. It is a customizable, cloud-based data pipeline made of a “quick start” set of tools that automatically process raw datasets in one place. The prototype standardizes, deduplicates, geocodes, and links the data. It also creates patient-level records to use for analysis.

Building blocks

“Building Blocks” make up the data pipeline. They are modular software services that accomplish one specific task, like geocoding or standardizing addresses. They can be reused by different programs or for different diseases or conditions. Single Building Blocks can be combined to create larger data processing and analysis pipelines.

What's next?

This pilot project is a first step in the data modernization journey. There is a lot more to try, discover, and understand as the public health community implements best practices at each step along the data journey from patient to public health and back.

During the next phase, the project team will

- Continue to assess STLT’s modernization challenges and involve them in the design process to ensure tools meet their needs over time
- Continue to test and improve the prototype so it can be used by different partners, tools, and systems
- Design more Building Blocks and, over time, stand up a marketplace where STLTs can access them

While the Building Blocks are being developed, there are a few things that STLT health departments can consider doing now to get ready for what’s next.

- Conduct a data and systems inventory to identify priority candidates for cloud migration, and begin the migration process
- Explore options for consolidated data hosting (e.g., data lakes) and use them for data consolidation and replacement of siloed systems
- Develop performance monitoring for their data ingestion pipeline to better identify problems and troubleshoot them in real time
- Maintain long-term access to raw, unprocessed data that has not yet been ingested into surveillance systems
- Increase the use of modern data processing and analytics tools that use open technologies (e.g., open source, standards, and architecture), such as
    - Open-source SQL-based relational database management systems
    - Data science and engineering languages like R and Python
    - Data processing, querying, and visualization tools (e.g., PowerBI, Tableau, Azure Synapse, and others)

Learn more

Read the complete summary of the pilot project in the report titled A Prototype of Modernized Public Health Infrastructure for All: Findings from a Virginia Pilot (November 2022).

Read Frequently Asked Questions: Data Pipeline Pilot Project.

This work is part of CDC's Data Modernization Initiative to modernize core data and surveillance infrastructure across the federal and state public health landscape.

Last Reviewed: February 17, 2023
Source: Centers for Disease Control and Prevention, Office of Public Health Data, Surveillance, and Technology