EXECUTIVE BRIEF

Creating a Cloud-based Data Pipeline to Enrich Public Health Data, Reduce Manual Processes, and Streamline Workflows

Findings from an LA County Pilot I December 2022 - December 2023

As a result of a pilot partnership between Los Angeles County (LAC), Centers for Disease Control and Prevention (CDC), and U.S. Digital Service (USDS), the LAC Department of Public Health is now using a cutting edge, modular, Data Integration Building Blocks (DIBBs) data pipeline to automatically process and enrich multiple data streams, including: COVID-19 electronic case reporting (eCR) files and electronic laboratory reports (ELR). This modern data pipeline was the output of a year-long development effort (Dec 2022-Dec 2023) that built on a previous pilot with the Virginia Department of Health (Jan-Sep 2022).

The DIBBs pipeline, which is composed of modular software components referred to as Building Blocks, has significantly reduced the time it takes for LAC's disease surveillance teams to receive and act upon public health data. This tool has the potential to notify case investigators of new cases of Hepatitis A approximately 19 hours faster than previous manual processes, from 20 hours to just 1 hour, or a 95% time savings. Furthermore, the pipeline can save LAC staff approximately 6 hours a week by collecting critical data elements for case investigation without manual intervention and reducing the need to manually run data processing scripts.

The LAC pilot clearly demonstrates how developing flexible, modular, and performant software that reduces manual work can get better, faster data to public health case investigators so they can take timely public health action.

PILOT GOALS

This pilot supports CDC's Public Health Data Strategy milestone of reducing manual time spent on preparing and harmonizing public health data by:

1/ Making eCR data available for LAC's public health case investigation

2/ Deploying an open-source, cloud-hosted data pipeline that can clean and transform multiple data streams

3/ Enabling LAC staff to own, operate, and maintain the pipeline

STRATEGY TO IMPROVE DATA WORKFLOW

Prior to the pilot, LAC could not use eCR data in a timely manner. While LAC had set up a data workflow to achieve tasks like analyzing the quality of sender data, this process was manual and inefficient, rendering the data less useful for case investigation by LAC disease teams. To streamline adoption, we developed a three-phased approach for deploying and operating our pipeline:

1/ Make LAC's existing workflow more efficient for upstream data analysts by establishing a single source of truth for eCR data

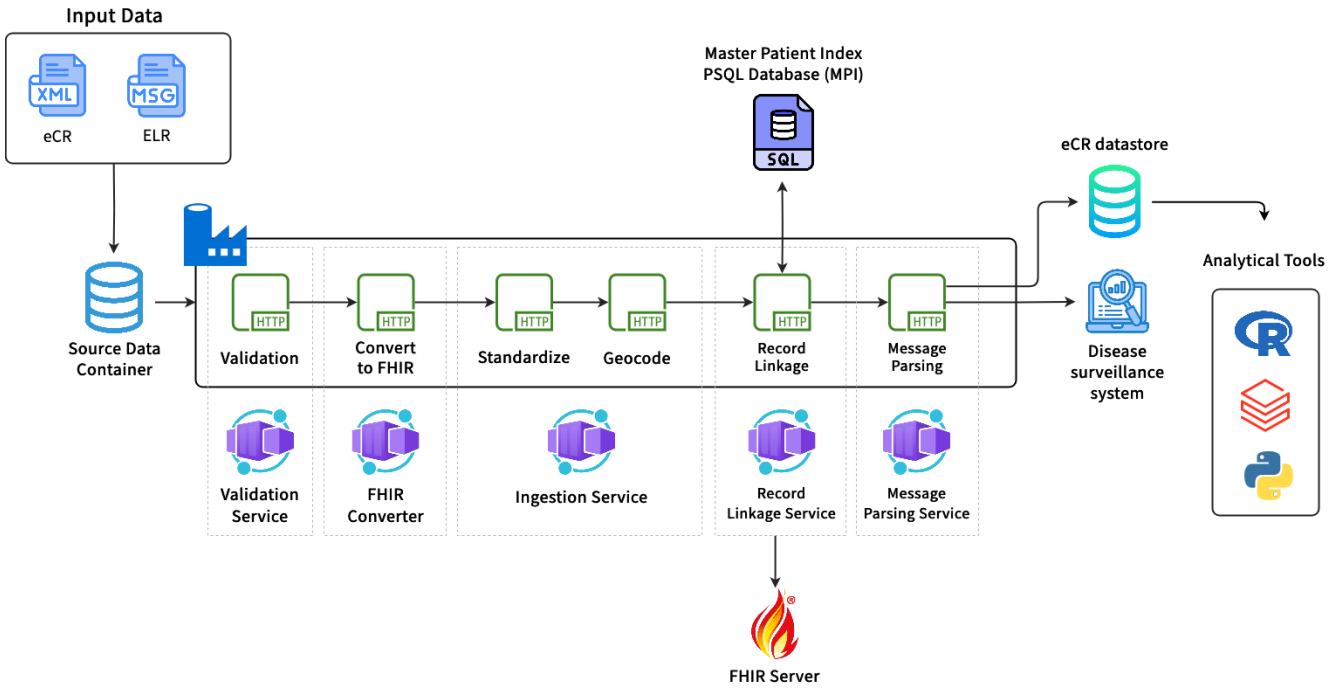
2/ Automate the delivery of analysis-ready data with fields relevant to downstream disease teams

3/ Enable LAC to operate the pipeline independently and expand use of eCR data to new program areas

DIBBS PIPELINE OVERVIEW

The DIBBS pipeline delivers real, production eCR data directly to LAC's case investigators. To enrich incoming public health data, the pipeline leverages a core set of five Building Blocks: **Validation, Conversion, Ingestion (includes Standardization and Geocoding), Record Linkage & Deduplication, and Message Parsing.**

The below architecture diagram visualizes how data moves through the DIBBs pipeline:



PIPELINE ADOPTION

The DIBBs team paired with LAC's eCR team to generate a data mart (e.g., tabular formatted dataset with relevant fields) for use downstream by LAC's Hepatitis team, which gives them access to processed eCR data for case investigation and analysis. Following user acceptance testing, the eCR team, on their own accord and without assistance from the DIBBs team, is developing a program-specific data mart for the Division of HIV and STD Prevention (DHSP). After DHSP, the eCR team plans to create another new data mart for the Community Outbreak Team (focused on viral respiratory pathogens). The eCR team can continue to leverage the DIBBs pipeline infrastructure to give additional disease teams access to processed eCR data.

PIPELINE IMPACT ON PUBLIC HEALTH

Key outcomes from using the DIBBs pipeline:

1/ Streamlined workflow, time savings, and greater data reliability – Cloud-hosted DIBBs pipeline simplifies the process for creating eCR datasets for downstream use, saving ~1 hour daily and mitigating the risk of system downtime.

2/ Faster onboarding of additional disease teams to eCR - eCR team can create program-specific data marts in a few hours rather than months, giving disease teams access to processed eCR data for case investigation and analysis.

3/ More timely eCR data for LAC's Hepatitis team - 95% faster (20 hours → 1 hour) for Hepatitis team to receive eCR data assuming an hourly refresh of the DIBBs pipeline's eCR data store.

4/ Reduced manual efforts to collect critical data elements - Estimated 12 minutes of time savings per Hepatitis A case, resulting in ~1 hour time savings a week.

5/ Increased data quality and completeness - Case investigators can quickly and easily identify positive Hepatitis A cases using additional data provided by the pipeline in an aggregated tabular format (vs. sifting through individual HTML files).

6/ Ability to process multiple data types - DIBBs pipeline also processed ELR data at a near 100% success rate without requiring significant modifications, enabling LAC disease teams to access and analyze cleaned ELR data downstream.

BEYOND THE LAC PILOT

Many of the public health data issues present in LAC's disease surveillance infrastructure (e.g., siloed, low-quality data) are common across public health jurisdictions. Through the LAC pilot, the DIBBs team gained insights on how to implement and translate our solutions to other jurisdictions and similar public health data challenges. Public health jurisdictions can learn more about our open-source DIBBs pipeline at https://cdcgov.github.io/dibbs-site/ and contact us directly at dibbs@cdc.gov.