February 28, 2023

A modern approach to standardizing and enriching public health data with modular building blocks in the cloud

Skylight

Who are we?

Skylight

Skylight is a digital consultancy using design, technology, and procurement to help agencies deliver better public services.

Dan Paseltiner — Data Engineer with ~8 years of experience building software to process and analyze data in the physical sciences, neuroscience, and public health (Maine CDC).

Brady Fausett — Data Engineer with ~20 years experience with healthcare data and interoperability (HL7v2, medical vocabularies, FHIR).

Our work in public health

Skylight

We've been a key partner of the CDC in designing the future of DMI:

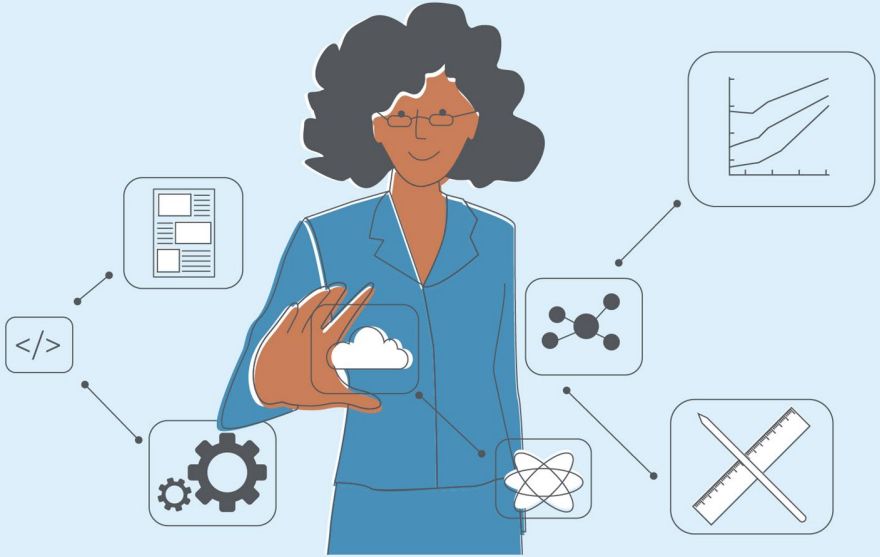
- We're at the forefront of building flexible, interoperable, and sustainable systems for public health.

- We built SimpleReport, a COVID-19 test result reporting tool that's processed over 7 million test results and counting.

- We are the engineers and researchers on CDC's PRIME Public Health Data Infrastructure (PHDI) project.

- We're also part of the CDC's FHIR Community of Practice.

What is Public Health Data Infrastructure (PHDI)?



Skylight

PHDI – The Problem

Working with public health data is challenging!

Volume

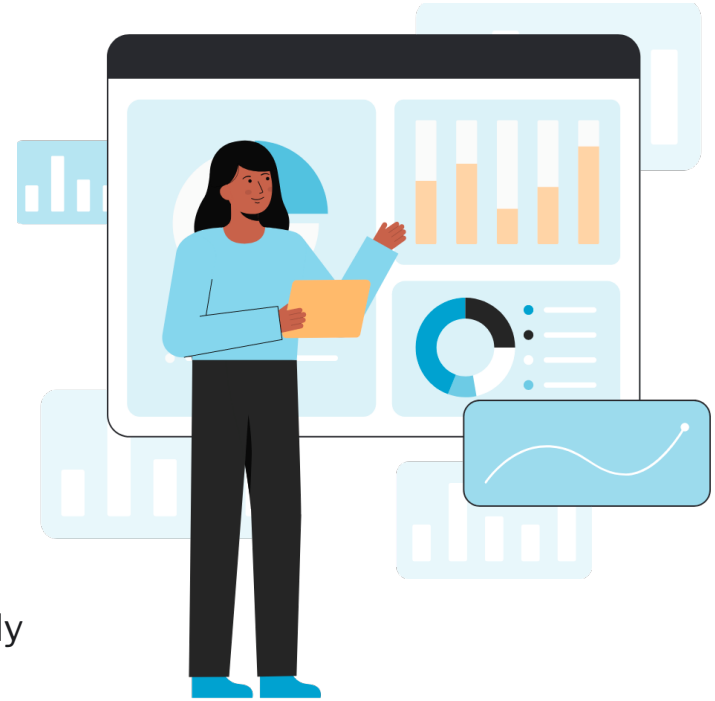
Velocity

Variety – many conditions, senders, and formats

Veracity – non-standard and potentially inaccurate

Value – low signal to noise ratio

Variability – incoming data and analytical needs change quickly



Skylight

PHDI – The Problem

Current Solutions are…

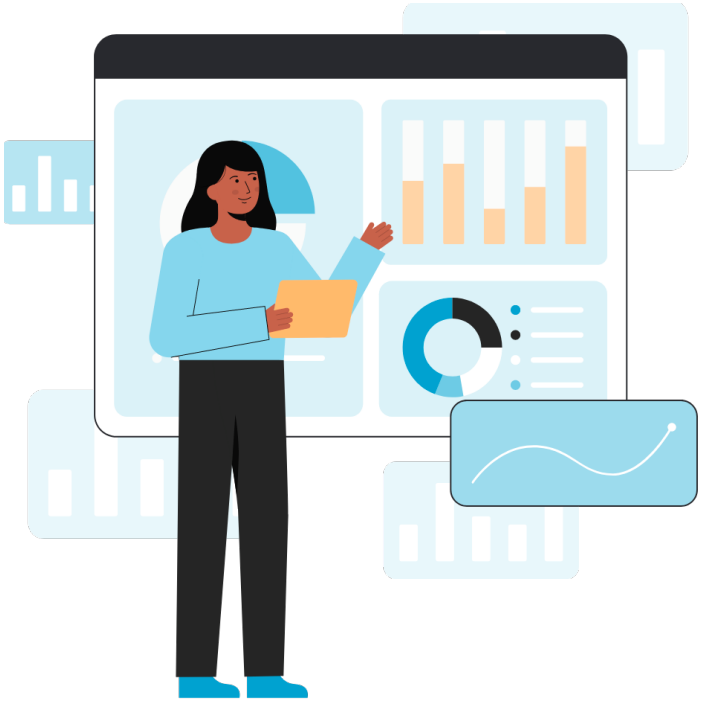
Bespoke

Sometimes manual

Inconsistent

Implemented throughout the data lifecycle

Rely on antiquated software and hardware



Skylight

PHDI – The Solution

Goal

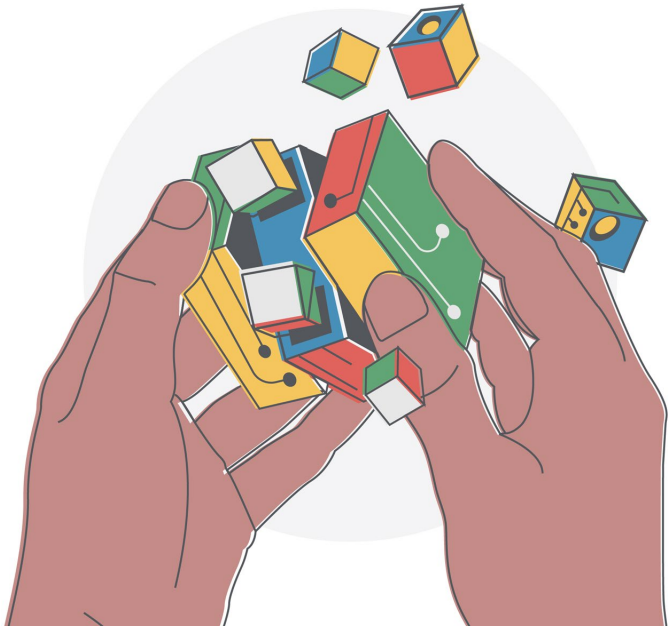
Improve data quality and reduce data cleaning workloads by providing analysis-ready data to downstream surveillance systems and other analytical and reporting applications.

Solution

Free, open-source modular Building Blocks for public health departments to build solutions that solve their data challenges.

Building Block examples:

- FHIR conversion
- Geocoding
- Tabulation of FHIR data



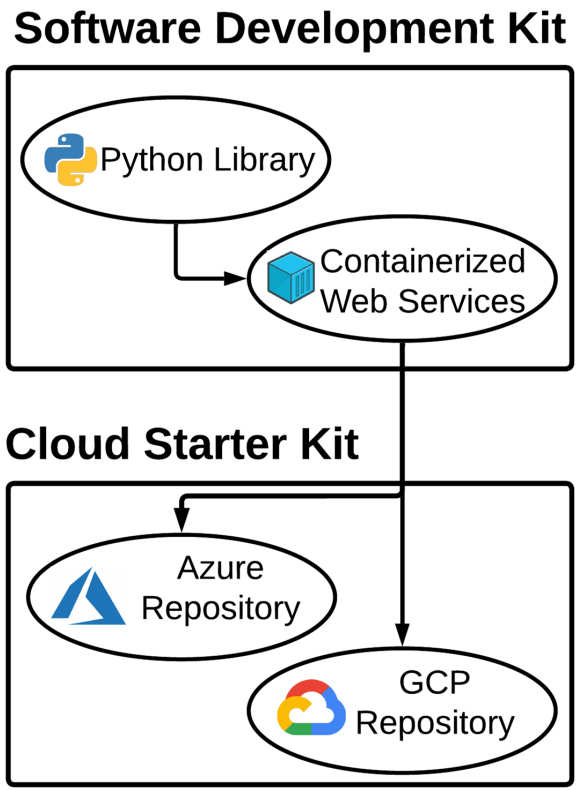
PHDI Products

Software Development Kit (SDK)

1. A Python library containing Building Block source code.
2. Containerized web services exposing Building Block functionality as HTTP endpoints.

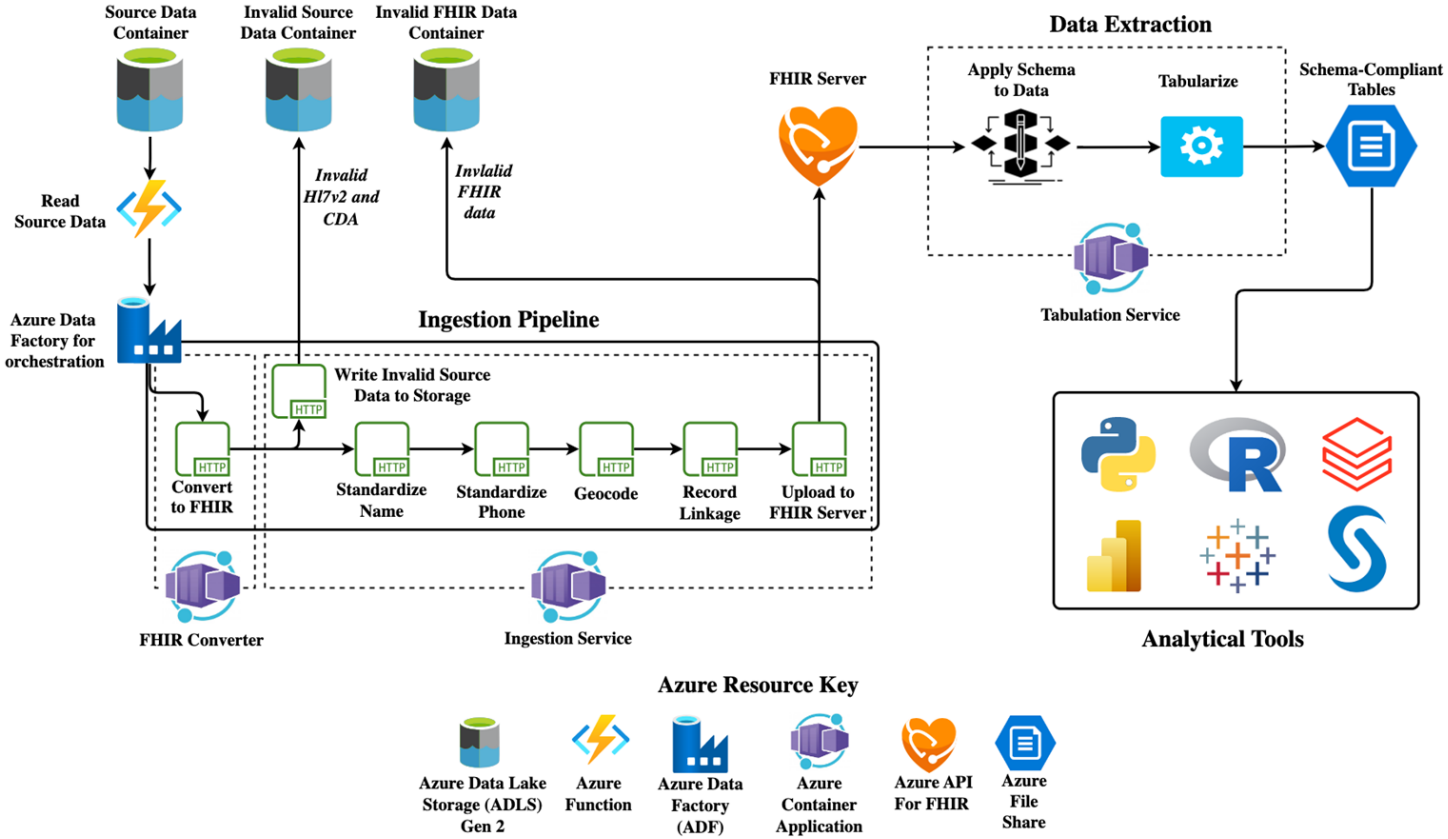
Cloud Starter Kit

- Repositories that implement a complete cloud-based pipeline composed on PHDI Building Blocks.
- User-friendly automated setup and deployment
- The start of a PH department's modern data infrastructure that.
- Aligns with CDC DMI and North Star goals



PHDI Starter Kit Architecture (Azure)

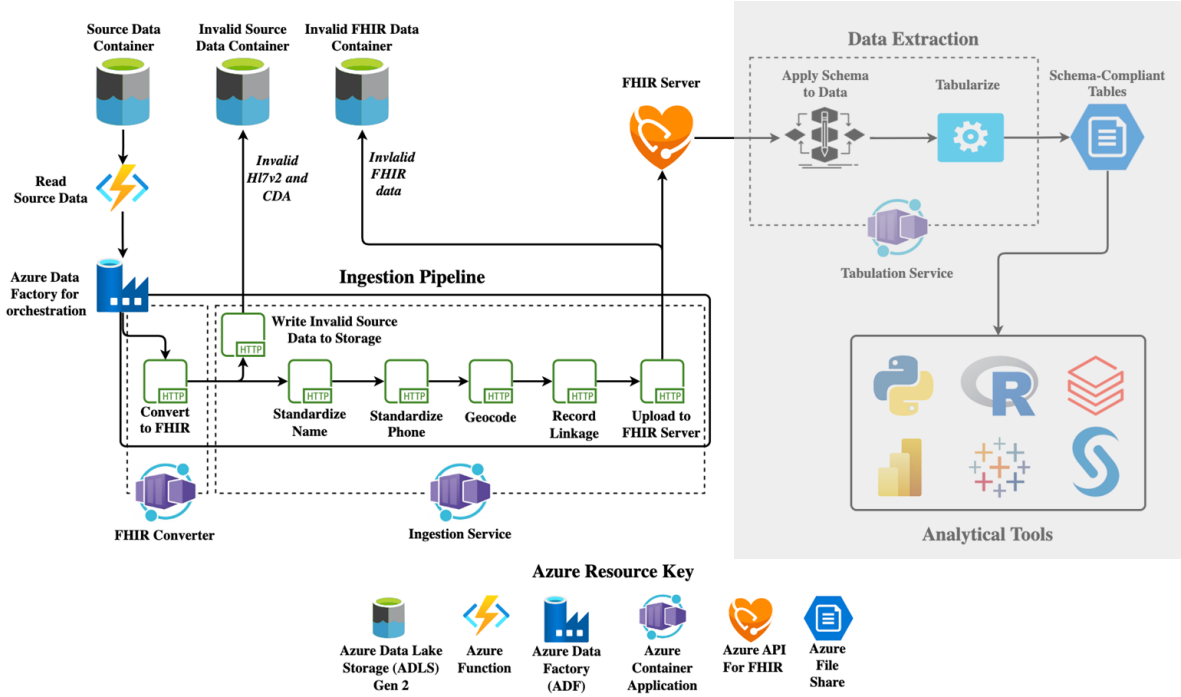
Skylight



Ingestion pipeline

- Convert to FHIR
- Clean and enrich
- Upload to FHIR Server
- Event-driven
- Serverless and scalable

Skylight



Conversion from HL7v2 to FHIR

Skylight

Input

Jan3e Doe | 2685 Sunrise Av Santa Rosa CA | (555) 604-7973 |

Output


```
"resourceType": "Patient",
"id": "c53f9ad8-34c1-ce05-c0f6-7a0ea7bd8483",
"name": [
  {
    "family": "Doe ",
    "given": [
      "Jan3e"
    ],
    "use": "official"
  }
],
  "address": [
  {
    "line": "2865 Sunrise Av",
    "city": "Santa Rosa",
    "state": "CA",
    "postalCode": ,
    "use": "home"
  }
],
"telecom": [
  {
    "system": "phone",
    "value": "(555) 604-7973",
    "use": "home"
  }
]
```

Name standardization

Skylight

Input


```
"resourceType": "Patient",
"id": "c53f9ad8-34c1-ce05-c0f6-7a0ea7bd8483",
"name": [
  {
    "family": "Doe",
    "given": [
      "Jan3e"
    ],
    "use": "official"
  }
],
  "address": [
  {
    "line": "2865 Sunrise Av",
    "city": "Santa Rosa",
    "state": "CA",
    "postalCode": ,
    "use": "home"
  }
],
"telecom": [
  {
    "system": "phone",
    "value": "(443)6047973",
    "use": "home"
  }
]
```


Output


```
"resourceType": "Patient",
"id": "c53f9ad8-34c1-ce05-c0f6-7a0ea7bd8483",
"name": [
  {
    "family": "DOE",
    "given": [
      "JANE"
    ],
    "use": "official",
    "extension": [
     {
      "url": "https://xlinux.nist.gov/dads/HTML/doubleMetaphone.html",
      "extension": [
        {
          "url": "familyName",
          "valueString": ["T", ""]
        },
        {
          "url": "givenName",
          "valueString": [["JN", "AN"]]
        },
       ],
     }
    ]
  }
],
  "address": [
```

Phone standardization

Skylight

Input


```
"resourceType": "Patient",
"id": "c53f9ad8-34c1-ce05-c0f6-7a0ea7bd8483",
"name": [
  {
    "family": "DOE",
    "given": [
      "JANE"
    ],
    "use": "official"
  }
],
  "address": [
  {
    "line": "2865 Sunrise Av",
    "city": "Santa Rosa",
    "state": "CA",
    "postalCode": ,
    "use": "home"
  }
],
"telecom": [
  {
    "system": "phone",
    "value": "(443) 604-7973",
    "use": "home"
  }
]
```


Output


```
"resourceType": "Patient",
"id": "c53f9ad8-34c1-ce05-c0f6-7a0ea7bd8483",
"name": [
  {
    "family": "DOE",
    "given": [
      "JANE"
    ],
    "use": "official"
  }
],
  "address": [
  {
    "line": "2865 Sunrise Av",
    "city": "Santa Rosa",
    "state": "CA",
    "postalCode": ,
    "use": "home"
  }
],
"telecom": [
  {
    "system": "phone",
    "value": "+14436047973",
    "use": "home"
  }
]
```

Address standardization and geocoding

Skylight

Input


```
"resourceType": "Patient",
"id": "c53f9ad8-34c1-ce05-c0f6-7a0ea7bd8483",
"name": [
  {
    "family": "DOE",
    "given": [
      "JANE"
    ],
    "use": "official"
  }
],
  "address": [
  {
    "line": "2865 Sunrise Av"
    "city": "Santa Rosa",
    "state": "CA",
    "postalCode":,
    "use": "home"
  }
],
"telecom": [
  {
    "system": "phone",
    "value": "+14436047973",
    "use": "home"
  }
]
```


Output


```
"resourceType": "Patient",
"id": "c53f9ad8-34c1-ce05-c0f6-7a0ea7bd8483",
"name": [
  {
    "family": "DOE",
    "given": [
      "JANE"
    ],
    "use": "official"
  }
],
  "address": [
  {
    "line": "2865 Sunrise Ave"
    "city": "Santa Rosa",
    "state": "CA",
    "postalCode": "95401",
    "latitude": 38.4404
    "longitude": 122.7141
    "use": "home"
  }
],
"telecom": [
  {
    "system": "phone",
    "value": "+14436047973",
    "use": "home"
  }
]
```

Record Linkage
& Patient De-duplication

Skylight

Applying record linkage to PH data streams

	eLR	eCR	VXU	Combined
% Reduction in # Patients	15%	29%	11%	19%

- Records linked using an identifier computed from name, address, and DOB.
- For **each individual stream**, a significant number of records were linked; **combining streams** yielded additional deduplication.
- **Converting to FHIR gives us identical fields across data streams** we can use to deduplicate and link records.

Demographic data recovery

- Expect increased prevalence of demographic data as a result of record linkage.
- 2-3% decrease in missing race/ethnicity fields in eLR data
- No recovery observed in eCR/VXU
- Combining streams yields addl ~1% recovery
- Small test dataset (18-day window) limits demographic linkage results

Recovery Example

Records pre-linkage:

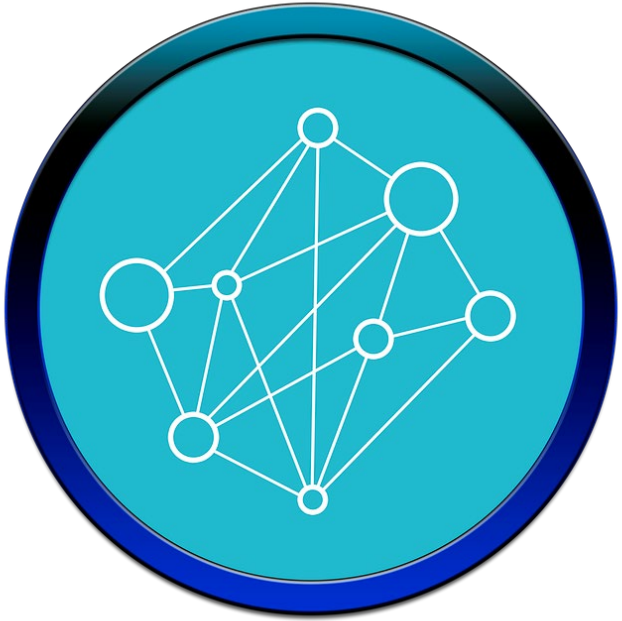
Jane Doe	Female	White	Non-hispanic
Jane Doe	Female		
John Doe	Male		Non-hispanic
John Doe		Black	Non-hispanic

Records post-linkage:

Jane Doe	Female	White	Non-hispanic
John Doe	Male	Black	Non-hispanic

Current/Future Record Linkage

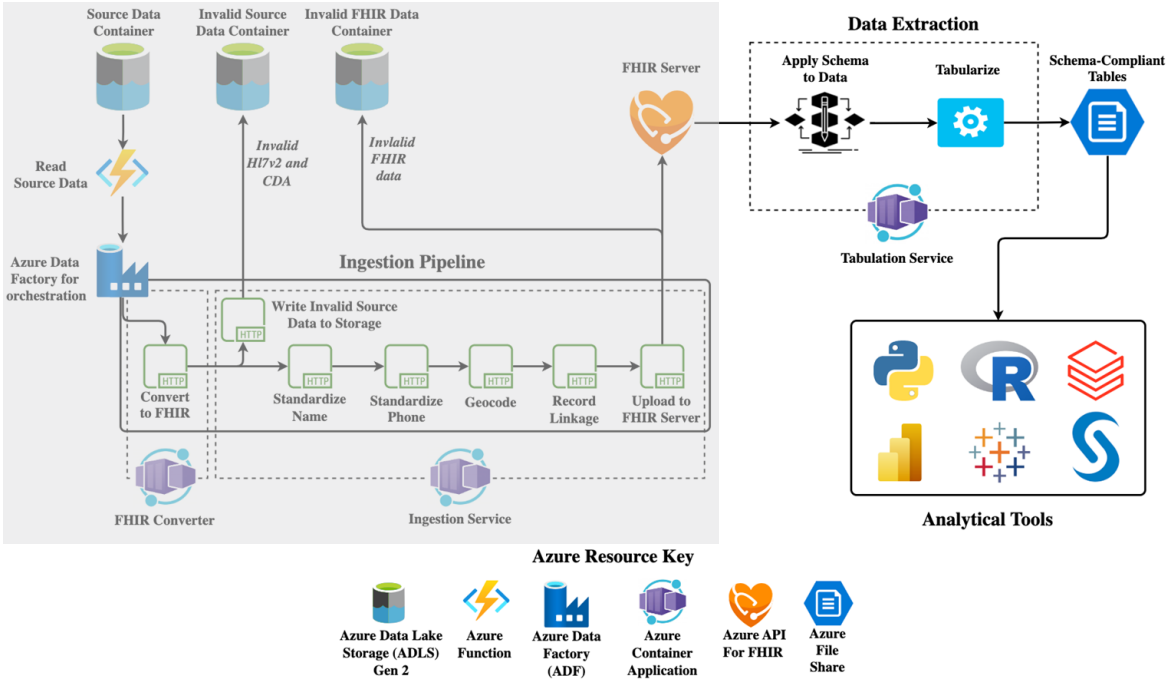
- Date of Birth Standardization
- Considering varied combinations of additional fields
  - Name, DOB, Address, Phone number
- Introducing probabilistic techniques
  - Levenshtein Edit Distance
  - Double Metaphone
  - Nickname resolution



Tabulation Service

- Returns data according to a user-defined schema
- Simple and flexible data access for epidemiologists and analysts

Skylight



Submit Schema to the tabulation service

1. Write a schema describing the desired data
   - tables and their columns

1. Submit the schema to the tabulation service

PATIENT_SUMMARY	
Patient ID	UUID
First Name	String
Last Name	String
DOB	Date
Phone Number	String

Data returned by the tabulation service

Patient ID	First Name	Last Name	DOB	Phone Number
1	CLARK	KENT	1938-02-29	+15555893245
2	HARRY	POTTER	1980-07-31	+15556924301
3	NANCY	DREW	1930-04-28	+15556256690

Persisted as flat files

- Parquet, CSV, SQLite
- Additional formats and direct database connections coming soon.

Custom schema generation + ease of use

Separation of schema generation from data storage

- Point-of-query schema application
  - High flexibility
- Customize fields for specific use cases

Declarative instead of imperative schema specification

- Epis and analysts simply state what data they want.
- No need to describe how to access the data.
- No Knowledge of FHIR, HI7v2, CDA, and database technologies is required.

Building Block Benefits

Standardize and enrich all message types consistently upon receipt

- eCR, ELR, VXU, and ADT

Flexible, modular, and interchangeable Building Blocks

- Containerized
- REST API

Event-driven serverless approach

- Data is processed in real-time
- Large batches are avoided
- Scale horizontally to meet demand
- Users to pay only for resources they use
- Infrastructure is managed by the cloud provider

Skylight

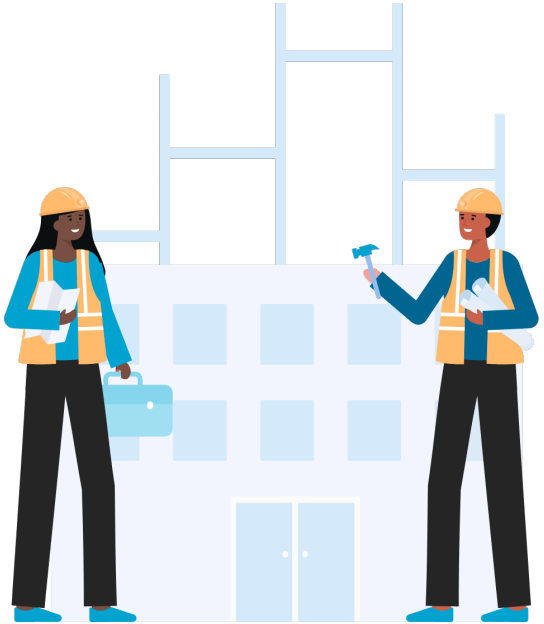
Coming soon!

Active Development

- Improved record linkage
- eCR validation
- Surveillance System Integration
  - Collaboration with NBS modernization

On Deck

- Incident level linkage (Case identification/de-duplication)
- HI7v2 (ELR, VXU, ADT) validation
- Terminology standardization (SNOMED, LOINC, etc…)



Skylight

Resources

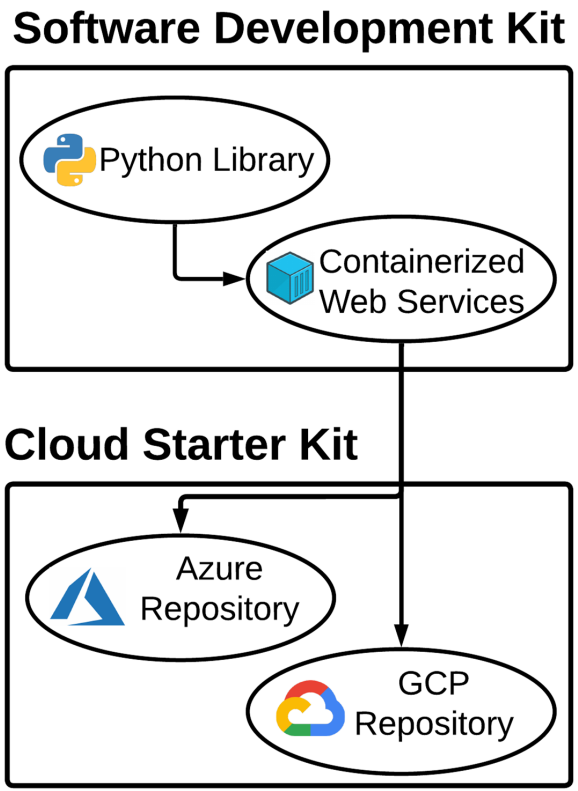
Open-source Building Block library

- Available as pypi package – `pip install phdi` today!
- GitHub repository
- Containerized FHIR Converter
- Containerized Ingestion Service

Starter Kit repositories for easy deployment

- Available for Azure and GCP

Learn more at: https://cdcgov.github.io/phdi-site/



Questions?

Skylight