

# Data Lakehouse for International Rescue Committee (IRC)

---

Samir Souidi  
JUNE 10-13, 2024



# About the Speaker



Samir Souidi is a Data Architect at the International Rescue Committee, leading the deployment of the Data Lakehouse architecture.

With over 20 years in IT, he specializes in data-driven solutions, technology optimization, and digital transformation.

Samir is also a travel foodie who enjoys interacting with local cultures.

# Table Contents

1. Overview of IRC's mission and data needs
2. Architecture Overview
3. Deployment challenges and solutions
4. Operation Efficiencies
5. Future Scope and Scalability



# International Rescue Committee (IRC)

IRC helps people affected by humanitarian crises—including the climate crisis—to survive, recover, and rebuild their lives.



## We serve

people whose lives have been upended by war, conflict and natural disasters



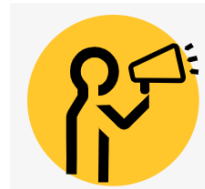
## We work

in countries where people don't have the support they need to recover from crisis



## We respond

within 72 hours, staying to help countries stabilize and people rebuild their lives



## We advocate

on behalf of people caught in crisis, encouraging governments to work smarter and do more

# IRC – Impact at a glance



**32.9 million+**  
People in countries reached by the IRC  
and partners in 2022

**1.4 million+**  
People reached with  
cash assistance

**\$109.7 M**  
Distributed in cash  
or voucher

**3,137**  
Supported health  
facilities

**222,278**  
Treated children  
under 5 for severe  
acute malnutrition

**807,853**  
Enrolled children  
and youth in  
learning programs

**8 million+**  
Provided primary  
health care  
consultations

# IRC Strategy

## IRC Strategy100

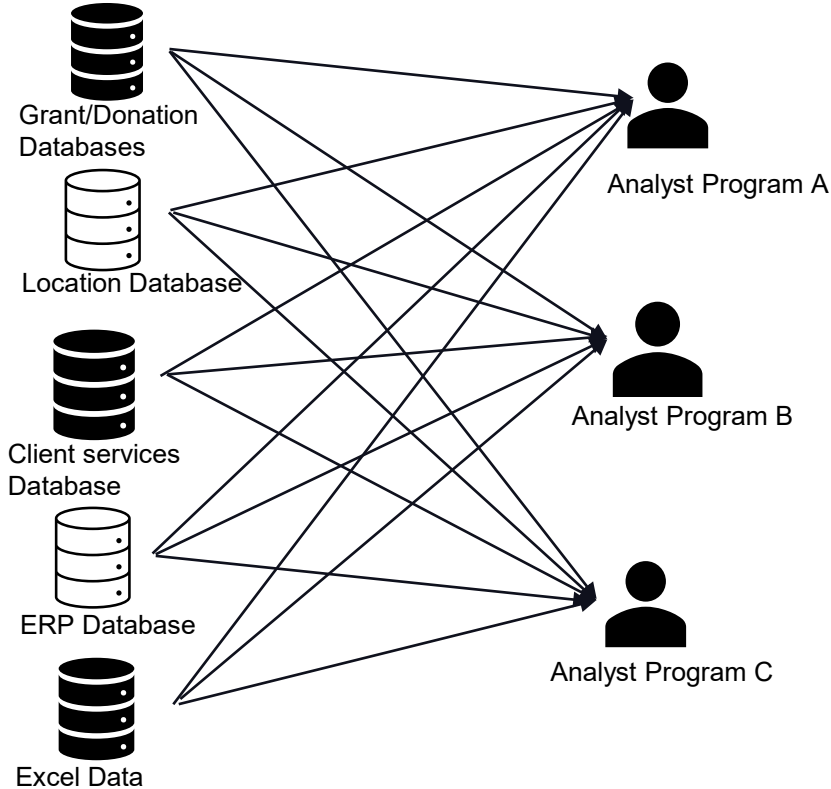
The goal of Strategy100 is to make our programs a **model** for the global humanitarian response. We aim to deliver **high-quality, cost-effective** programs—on our own and with **local partners**. We will combine research on programs that work best with **insights driven** by those we serve, to reshape the way the world helps those in need. With this we want to make **empowerment** and lasting change the norm.

## IRC Data Ambition

- Strengthen and enable systematic use of data across the organization.
- Enhance skills and culture of data use.
- Improve data collection and visualization.
- Utilize advanced data management tools.

# IRC – Data Analytics Process

## Previous Architecture

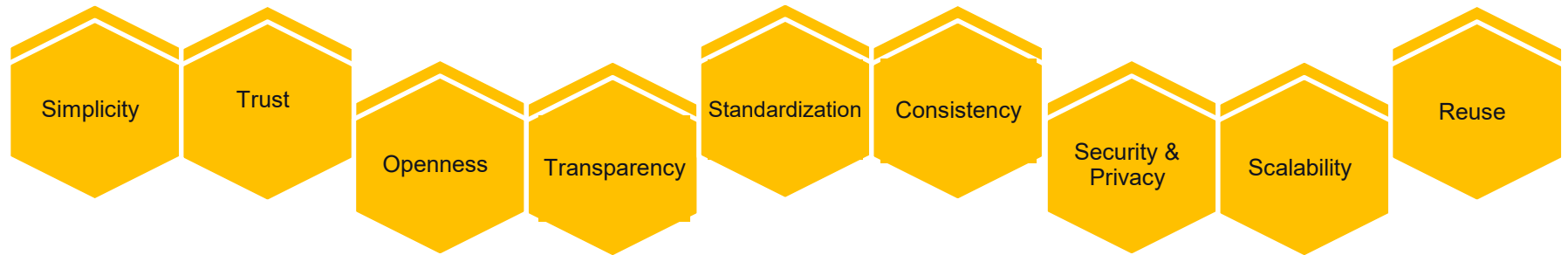


## Challenges

- Silos in databases and data models.
- Reporting data limited to function/process-defined data mapping and integration.
- Time-consuming data transformation to fit current department/unit needs.
- Limited capacity for cross-sharing knowledge of data model build-up.
- Data quality issues.
- Current technology stack limitations.
- Lengthy process to access data.
- Limited data catalog across IRC databases.
- No Enterprise IRC Data Warehouse.

# IRC Lakehouse

## Data Architecture Principles

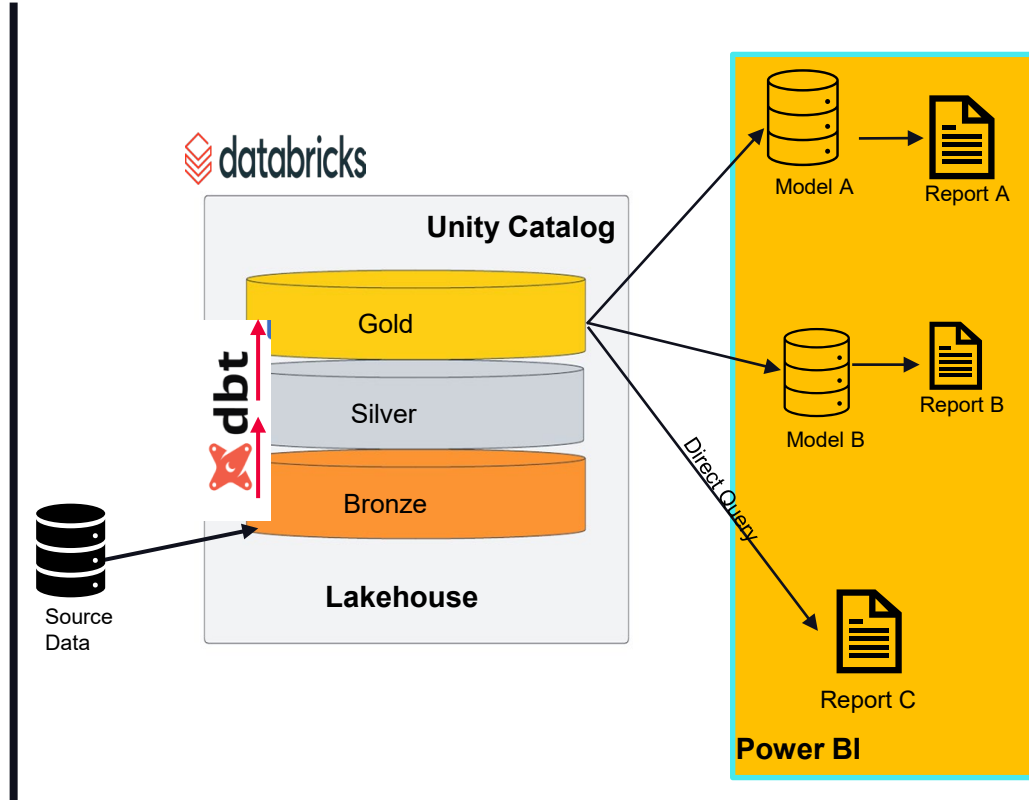
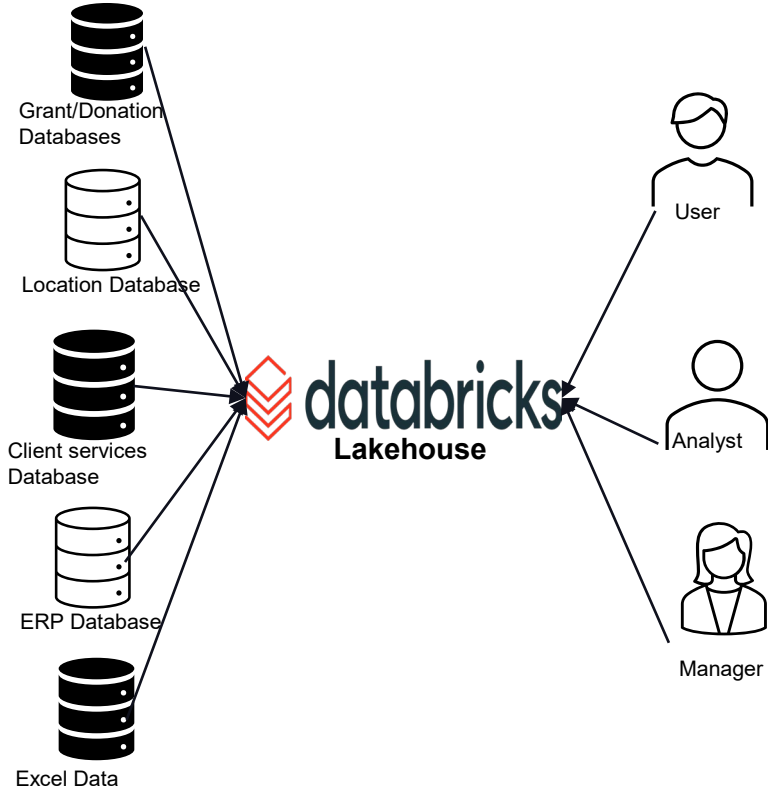


The Enterprise Data Lakehouse will support better decision-making, enhance operational efficiency, and help IRC achieve its Data S100 goals more effectively.

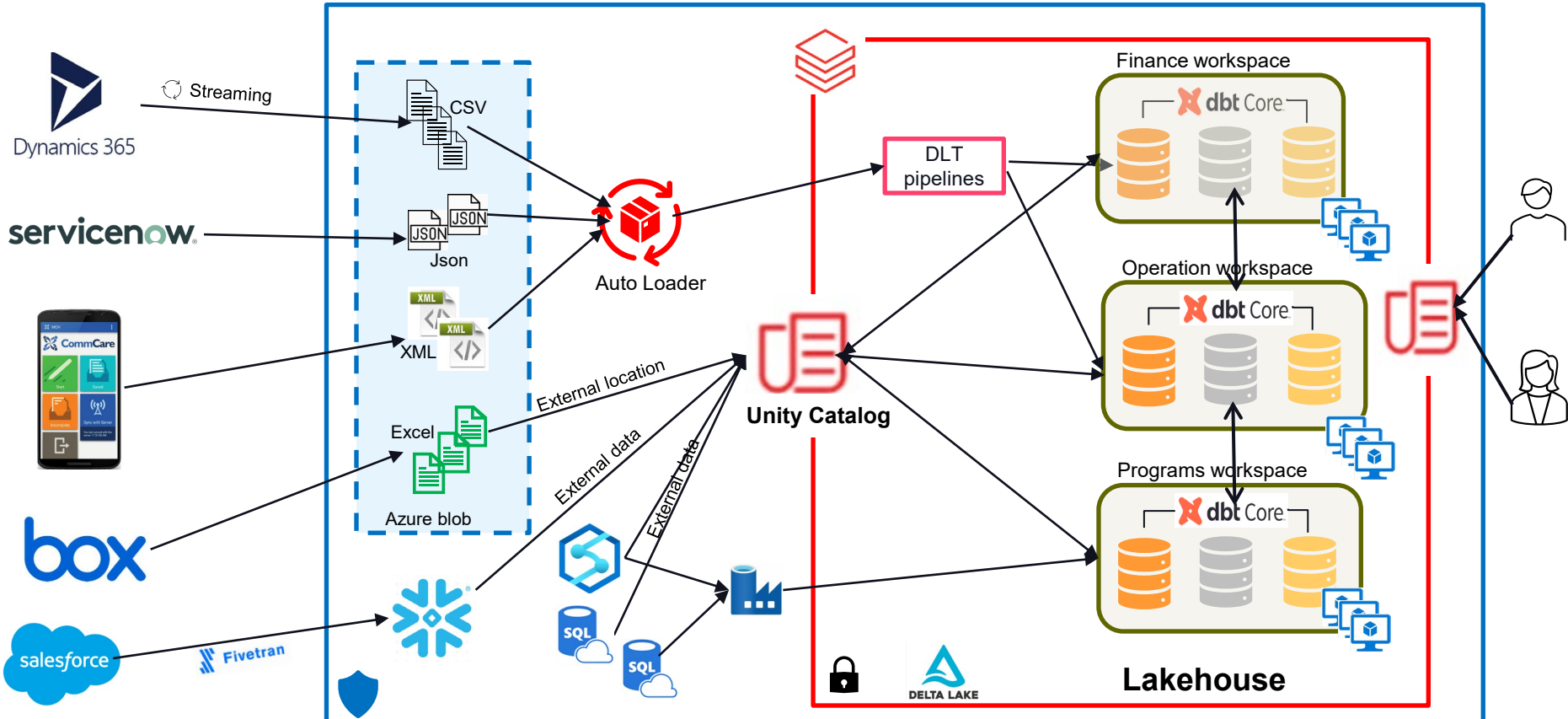


# IRC – Data Analytics Process

## Current Architecture: Lakehouse and Medallion Architecture



# IRC - Lakehouse Architecture



# Deployment Finance Lakehouse

## Dynamcis365 F&O to Lakehouse with streaming DLT + dbt

The image displays a Databricks workspace interface. On the left, a 'Catalog' sidebar shows a tree view of folders including 'In my org' and 'Shared'. The main area contains a code editor with a Python script for a streaming DLT pipeline. The script defines functions for reading data from cloud files, ingesting CSV data, and applying changes to a target table. On the right, a dependency graph shows 'Integra\_job' (Pipeline: integra\_dlt) pointing to 'finance\_accounting\_dbt' (3 dbt commands, dbt\_CUI). Below this, a 'Workflows' view for 'Delta Live Tables' shows a graph of streaming tables: 'temp\_v\_ledgerjour...', 'ledgerjournaltrans...', 'temp\_v\_ledgerjour...', 'logisticsaddressco...', and 'temp\_v\_logisticsad...'. Each table node includes completion status and metrics.

```
def get_cdc_data(table_, file_location, schema_):
    @dlt.view(name=f"temp_v_{table_}")
    def ingest_csv_data():
        df= spark.readstream.format("cloudfiles")\
            .option("cloudfiles.format", "csv")\
            .option("header", "false")\
            .option("delimiter", ",")\
            .option("escape", "\\")\
            .option("newline", "\\n")\
            .option("multiline", "True")\
            .schema(eval(schema_))\
            .load(file_location).withColumn("_SysRowId", col("RECID"))\
            .withColumnRenamed("Start_LSN", "LSN").fillna('').fillna(0)\
            .withColumn("default_date", lit(datetime.datetime(1900,1,1)))

        for i in [f.name for f in df.schema.fields if isinstance(f.dataType, TimestampType)]:
            df = df.withColumn(i, when((col(i).isNull()), col("default_date")).otherwise(col(i)))
        df =df.drop(col("default_date"))

    return df

dlt.create_streaming_table(f"{table_}")

dlt.apply_changes(
    target=f"{table_}",
    source=f"temp_v_{table_}",
    keys = ["RECID"],
    sequence_by = sf.col("LSN"),
    apply_as_deletes = sf.expr("DML_Action = 'DELETE'"),
    except_column_list = ["DML_Action", "Update_Mask", "Seq_Val", "End_LSN"]
)
```

# Deployment Azure log dlt

## Azure Databricks audit log and Monitoring

The screenshot displays the Databricks workspace interface for a pipeline named "Azure\_log\_dlt". The pipeline is shown as completed on 5/27/2024 at 5:00:21 AM. The task graph includes several streaming tables and alert tasks:

- Streaming tables: insights\_logs\_inte..., silver\_pipeline\_runs, insights\_logs\_stor..., insights\_logs\_stor..., storage\_clean, insights\_logs\_stor..., insights\_logs\_thre...
- Tasks: new\_ip\_alert, Add\_new\_ip\_address, req\_size\_alert, and display\_req\_size.

An inset window shows the task graph for "Azure\_Log\_DLT", highlighting the sequence: new\_ip\_alert (New IP Address Alert) → Add\_new\_ip\_address (Monitor Scripts/takehouse\_log\_script) → req\_size\_alert (Storage Req Size Threshold Alert) → display\_req\_size (Monitor Scripts/display\_highreq\_size). A "+ Add task" button is visible below the graph.

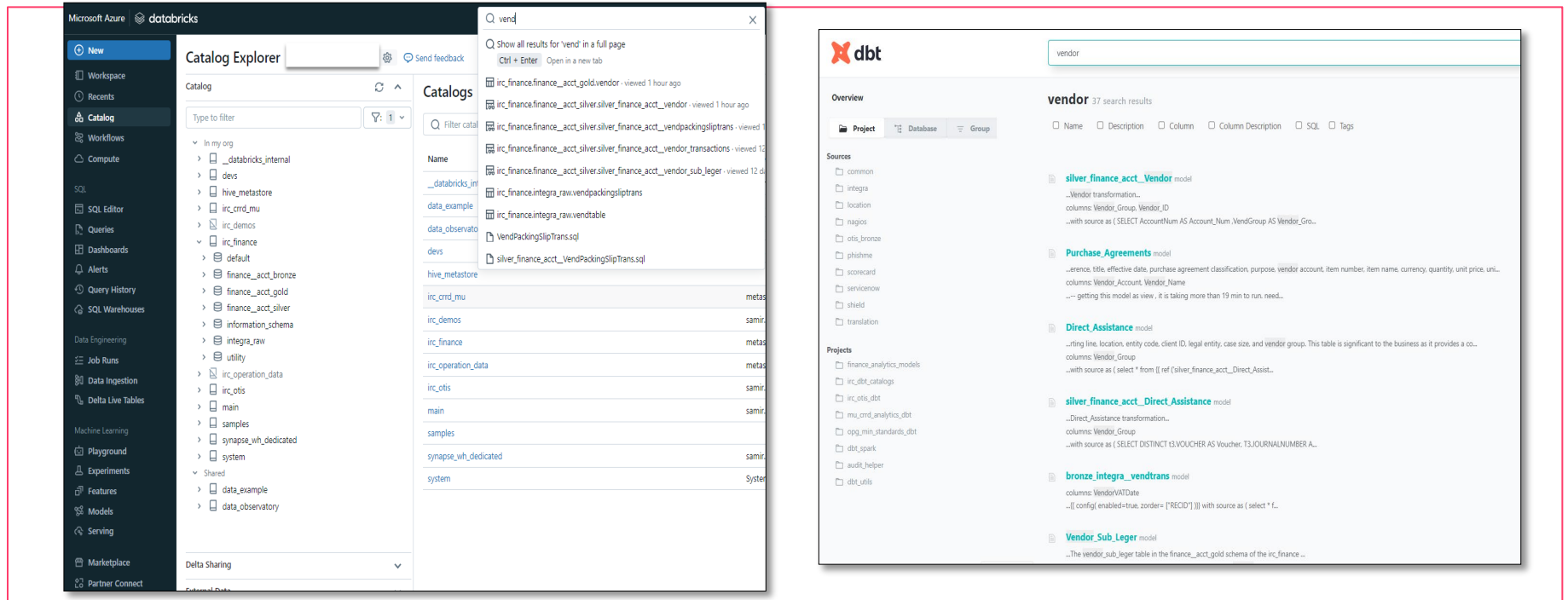
Another inset window shows a SQL query and its results:

```
1 select format_number(count (*), '000,000') as total_recod_number
2 from storage_clean
3
```

Raw results	total_recod_number
1	6,362,869,932

# Deployment Unity Catalog

One catalog source – build trust and transparency

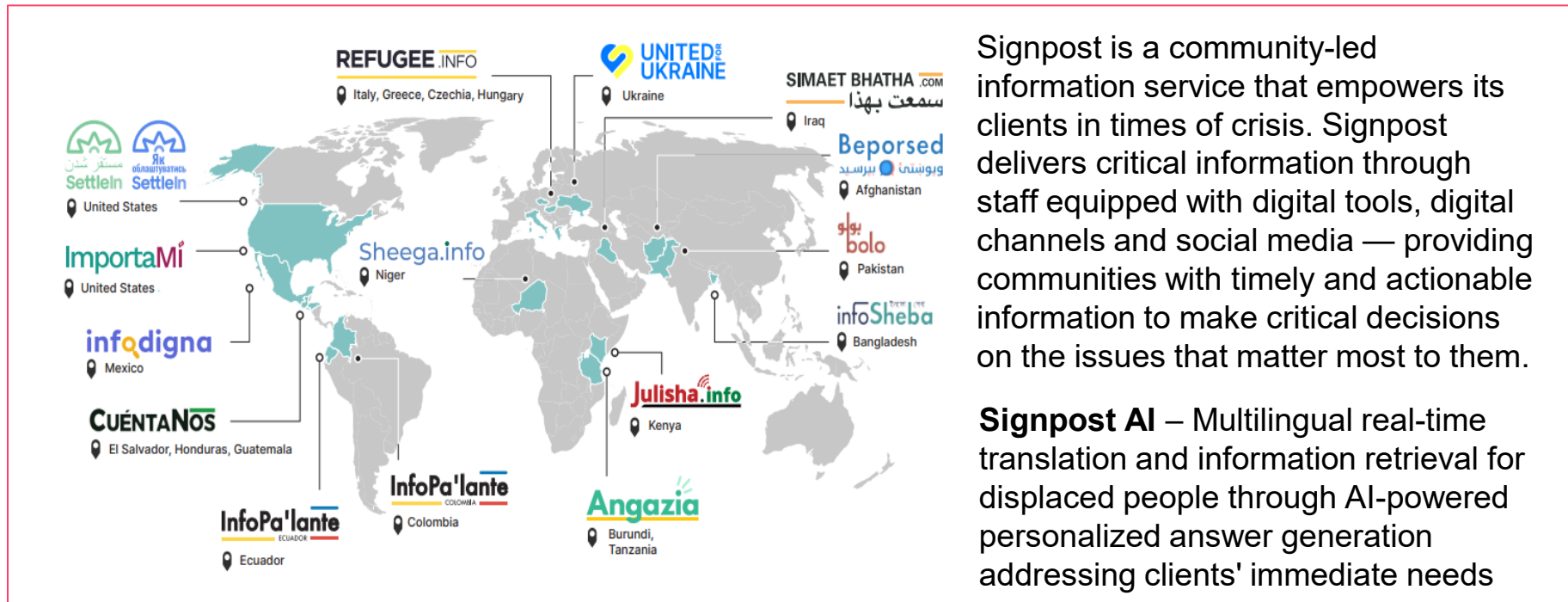


# Benefits of the Data Lakehouse Implementation

- . Near real-time access to data for data analysis.
- . A simpler process to access data
- . Reduce the data redundancy
- . Cost efficiency and scalable based on the demand, not one size fits all
- . Allow decision-makers to be better informed and on time.
- . Monitoring projects and grants at near-closed
- . One data dictionary

# Future Project

## Signpost AI – information access



A world map with callouts to various countries, each accompanied by a logo and the name of a Signpost AI service. The services are distributed across North America, South America, Europe, Africa, and Asia. The logos include: Settlin (United States), ImportaMi (United States), infodigna (Mexico), CUÉNTANOS (El Salvador, Honduras, Guatemala), InfoPa'lante (Ecuador), InfoPa'lante (Colombia), REFUGEE.INFO (Italy, Greece, Czechia, Hungary), UNITED OF UKRAINE (Ukraine), Sheega.info (Niger), Julisha.info (Kenya), Angazia (Burundi, Tanzania), SIMAET BHATHA.COM (Iraq), Beporsed (Afghanistan), bolo (Pakistan), and infoSheba (Bangladesh).

Signpost is a community-led information service that empowers its clients in times of crisis. Signpost delivers critical information through staff equipped with digital tools, digital channels and social media — providing communities with timely and actionable information to make critical decisions on the issues that matter most to them.

**Signpost AI** – Multilingual real-time translation and information retrieval for displaced people through AI-powered personalized answer generation addressing clients' immediate needs



# Q&A