# **RTV Data Engineer Technical Assessment**

**About Us**

At Raising The Village (RTV), we are dedicated to eradicating ultra-poverty in Sub-Saharan Africa. As a dynamic, rapidly growing international development organization, we’ve assembled a team of over 250 passionate individuals in Uganda, alongside an additional 17 professionals in North America and 15 in Rwanda. Together, we are committed to elevating communities out of ultra-poverty by implementing innovative solutions and leveraging advanced data analytics to drive impact.

To date, our holistic approach has positively impacted over 1 million lives since 2012, and we’re poised to achieve even greater milestones, aiming to assist 1 million individuals annually by 2027. Our growth and success are fueled by the invaluable support of global partners who share our vision of sustainable change. Learn more about our impactful programs at [www.raisingthevillage.org](http://www.raisingthevillage.org)

The **VENN department** is the data and technology backbone of our organization, connecting advanced analytics, and custom software tools with field implementation to ensure data-informed decision-making at every level.

## **Assessment Overview**

* **Time: 6Hours**
* **Points:60**
* **For any follow-up questions please contact-0757005504/0774128257**

This assessment evaluates your proficiency in designing and implementing end-to-end data engineering solutions using RTV household survey data. You will demonstrate expertise in data architecture, ETL/ELT processes, data warehousing, and analytical visualization.

## **Submission Instructions**

### **Required Deliverables**

* **Source Code**: Complete, executable codebase with clear README
* **Documentation**: Architecture diagrams, setup instructions, and design rationale
* **Dashboard**: simple dashboard with insights

### **Submission Format**

* GitHub repository with organized folder structure
* Provide clear instructions for running your solution

## **Technical Challenge**

**Primary Objective**: Design and build a complete data pipeline and analytical dashboard that enables RTV to track poverty-related household progress metrics over time using our sample simulated cohort data.

### **Dataset Overview**

The artifacts (Data, data collection tool ) to use are zipped in a folder “combined\_data”.

For simplicity, you will work with the cohort spanning 3 survey cycles:

* Baseline Survey
* Year 1 Follow-up
* Year 2 Follow-up

The corresponding collection tools for the datasets are labeled as follows:

* ahs\_2021\_baseline.html
* ahs\_2021\_year1.html
* ahs\_2021\_year2.html

## **Deliverables**

### **Data Pipeline Architecture**

Design and implement a modern data pipeline solution that addresses longitudinal household survey data challenges:

**Core Components**:

* **Data Lake Foundation**: Object storage as raw data layer (MinIO, AWS S3, GCS, Azure Blob)
* **Data Ingestion**: Strategy for CSV files with 1,000-5,000+ variables
* **ELT vs ETL**: Justify your transformation strategy choice
* **Schema Management**: Handle evolving survey structures and variable changes
* **Data Quality**: Validation, profiling, and cleansing methodologies
* **Orchestration**: Workflow management and dependency handling
* **Monitoring & Observability**: Data lineage, pipeline health, error handling

**Technology Options**:

* Local: Python/pandas, Apache Spark, dbt Core, Airflow, PostgreSQL
* Commercial/Cloud: Snowflake, BigQuery, Databricks, Fivetran, DBT Cloud, Prefect
* Hybrid: Mix local and cloud components

**Deliverable**: Executable code, architecture diagrams, design document explaining technology choices

### **2. Data Warehouse & Transformation Strategy**

Implement a warehouse solution optimized for analytical workloads on longitudinal survey data:

**Key Considerations**:

* **Warehouse Selection**: Choose and justify technology (PostgreSQL, Snowflake, BigQuery, Databricks)
* **ELT Implementation**: Design transformation layer within warehouse
* **Schema Design**: Structure for efficient querying of high-dimensional survey data
* **Performance Optimization**: Partitioning, clustering, indexing for large datasets
* **Data Modeling**: Dimensional modeling, data vault, or other suitable approaches
* **Variable Cataloging**: Metadata management for survey variables and evolution

**Transformation Challenges**:

* Multi-level geographic hierarchy management
* Seasonal/temporal data patterns
* Variable normalization across years with changing names/definitions
* Derived metrics and business logic implementation

**Deliverable**: Complete warehouse implementation with transformation logic, schema documentation, performance analysis

### **3. Analytical Dashboard**

Build a dashboard enabling stakeholders to explore insights from the data warehouse. Focus on visualizations valuable for RTV's poverty-fighting mission.

**Visualization Options**:

* Local: Streamlit, Plotly Dash, Jupyter notebooks, Grafana
* Commercial: Tableau, Power BI, Looker, Hex, Observable
* Custom: web applications

**Deliverable**: Simple dashboard with with user documentation

### **4. Technical Documentation**

**Required**:

* README with setup and execution instructions
* Code documentation and inline comments
* Data quality assessment and validation results
* Performance considerations and scalability notes
* Automated testing frameworks

### **Infrastructure Options**

Choose the approach that best demonstrates your skills:

**Local Development**

* Containerized stack using Docker Compose for easy evaluation
* Object storage simulation with MinIO or local filesystem
* PostgreSQL, DuckDB, or ClickHouse as warehouse
* Include setup scripts and clear documentation

**Cloud/Commercial**

* Managed services (Snowflake, BigQuery, Databricks, etc.)
* Document costs and provide screenshots/demos for evaluation
* Include Infrastructure as Code (Terraform/CloudFormation) if applicable

**Hybrid**

* Combine local orchestration with cloud warehouse
* Balance cost-effectiveness with feature requirements