Data Processing Pipeline for Field-Collected Data

Submitted by Dave Henry (eHealth Africa) on January 18, 2018 - 5:43am
Last revised by Web Producer on June 21, 2018 - 3:09pm.
Proposal Status: In Review

Background

eHealth Africa (eHA) is a non-governmental organization focused on improving health systems with core technical expertise driven by technology in Health Delivery Systems, Public Health Emergency Management Systems, Disease Surveillance Systems, Laboratory & Diagnostics systems, Nutrition & Food Security Systems and in sustaining program interventions.

We work primarily in Nigeria, Sierra Leone and Liberia with resource capacity and strategic partnerships on programs/projects in Cameroon, Chad, Democratic Republic of Congo, Mali, Niger, South Sudan, and Somalia.

Executive Summary

The initial challenges of mobile field data collection – offline operation and reliable transport of media-rich, programmable forms – have been solved. Tools such as ODK make it possible for thousands of organizations to conduct field surveys, to gain insight into population health and to plan for effective interventions. eHealth Africa's proposal seeks to improve upon Open Data Kit (ODK 1) by addressing the challenges that occur after initial collection:

- Field data is not readily available for analysis, it typically needs to be downloaded and staged for access by end-users
- Organizations that need to officially publish field data require a workflow and automation infrastructure
- Field data is not registered and tracked, which limits its ability to be reliably updated over time
- Field data is not strongly "typed", making it difficult to ensure conformance for interoperability use cases

eHA has been working on a solution to these problems for the past year and has performed initial deployments in DRC, Nigeria, Sierra Leone and Chad. The software is now mature enough to share with the community at large and we seek collaborators and additional funding to accelerate the roadmap. The outcome of this collaboration will provide the following benefits:

- The ability for researchers and stakeholders to access field data in real time from a secure data portal (e.g. CKAN, HDX)
- Support for "revise and release" workflows using a 3rd-party open-source BPMN-based workflow engine. (Canonical use case: MoH publication of master facility lists)
- The ability to uniquely identify and version each record that enters the system, enabling applications to update and submit changes without loss of provenance. (Canonical use case: longitudinal updates to patient health history records)
- Support for explicit record schemas, ensuring that data which is collected in the field is correctly structured for use in standards-based exchanges (Canonical use case: collection of health facility data using OpenHIE / FHIR facility profiles)

Consortium Team

The eHealth Africa team currently consists of a development lead, a systems architect, a product manager, a DevOps engineer and three senior developers. Collectively the team is skilled in solution architecture, UX design, software development, DevOps, product management, open source community development and vendor relationship management. ***We are seeking collaborators who are interested in a co-development role as well as field-testing of specific use cases.*** We intend to release the source code in May 2018 under an Apache 2.0 license.

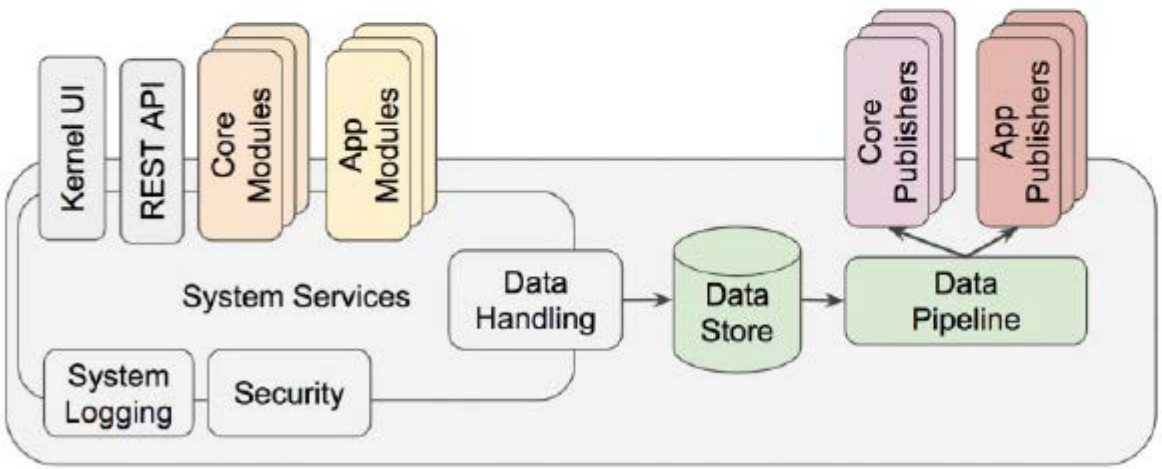
The two principals for this proposal are Dave Henry (https://www.linkedin.com/in/dave-henry-968431/) and Adam Butler (https://www.linkedin.com/in/adamvert/) . Dave is Director of Global Health Informatics at eHA and Adam is Technical Team Lead. Adam

has recently been selected as a member of the ODK 1 Technical Steering Committee (TSC).

Project Description

As part of an effort to develop a pre-integrated software stack for GIS field data collection, eHA has improved how ODK 1 is used in data processing pipelines. We have created a solution that associates ODK 1 forms with schemas and extracts entities from ODK Collect form submissions. Specifically, data is serialized into Avro with an accompanying schema. An open source project called Salad (Semantic Annotations for Linked Avro Data) is used to link the Avro schemas via JSON-LD. The extracted data is semantically typed as it moves through the transformation and enrichment process. This offers several benefits:

- Since the schema is kept separate from the form definition, changes to the form definition do not require creation of a new data set
- Multiple forms can use the same schema and "entity store"
- Hierarchical data can be automatically normalized as tables in a relational database with pre-assigned primary / foreign keys
- The availability of unique keys provides the basis for deterministically updating data that has been distributed to mobile clients. This sets the stage for the "offline record update" scenario that is important to many organizations (e.g. longitudinal patient record management)
- External schema repositories (e.g. Schema.org) can be used to support collaboration between organizations and across projects
- The core solution stack consists of Python, Django and PostgreSQL. A companion module uses Kafka to stage data in a topic-oriented repository where "publisher" applications can subscribe to changes and forward blended / reformatted data to target destinations. The architecture is functionally represented as follows:



The architecture enables developers to add new capabilities via plugins on both the data collection and the data publishing side. Core plugins are generally useful and may be shared openly via a central repository. App plugins are application specific capabilities for building bespoke solutions. Both types of plugins interact with the system via well-defined REST APIs.

The stack will be "Dockerized" and available on DockerHub.