Summary Report of P20WIN Pre-Production Testing

Categories of Test Reports

The PATH software application is made up of three basic components: A Remote component installed at a data owner’s site, a Main component and the Probabilistic Integration component, both installed at a central networked location. The Pre-production Testing carried out by UCHC tested the features of these three Modules as well as end-to-end testing of the entire application. These four categories of Test Reports are available for review.

- **The Metadata Editor (MDE)** is part of the Remote. It allows the Data Steward to specify the data schema for the dataset(s) that will be integrated and shared.
- **The User Interface (UI)** is part of the Main. It supports all User Roles (Data Consumer, Data Steward and System Administrator) to carry out their functions of assigning and using logins and passwords, defining data, defining queries, reviewing and approving queries, executing queries and downloading data resulting from queries.
- **The Probabilistic Integration (Pi)** is also part of the Main. It builds a cluster of discrete data records from multiple sources into one cluster of information about one person.
- **End-to-End** testing was carried out by:
  - Defining the data sets using the MDE
  - Initializing the system that would test the correct performance of the Remote ETL module and would cause Pi to cluster the records
  - Executing queries that would test the correct functioning of the Main Query Engine.

Test Results

**MDE and UI** - Tests plans for the MDE and UI were created by the P20WIN Project Leadership Team (PLT) to verify that all the requirements in the Memorandum of Agreement were satisfied. UCHC used these test plans. The full report of UI and MDE testing maps each test to an MOA Requirement so that completion of the MOA was also validated along with the correct execution of the feature. These full tests reports are posted for public review.

Results of the UI Test - 83 tests were executed, 82 passed with one partial pass. (The time stamp on the Audit Log Entry for Query Creation was the Query Approval time instead.) All MDE tests passed completely.

**Pi** - Testing the Pi module within the P20WIN environment was challenging for 3 reasons.

1. Pi uses a probabilistic integration method that is based on configurable parameters set by UCHC. This enables the matching algorithm to be tuned, so that the best possible matching can be achieved for the particular characteristics of the data being integrated. The mechanism for setting the Pi matching parameters is an iterative process of set-test-review with a manageable sample size of representative data.
2. The PATH application delivers a de-identified dataset of Fact records with all the Personal Identifiable Information (PII) excluded from the query result. This makes a review of how well Pi matched PII information impossible within the PATH application. Therefore, UCHC asked their vendor, OpenClinica to develop a standalone Desktop Pi Utility which performs the same function as other software in marketplace. Like Data Ladder, Desktop Pi does not deliver an anonymized dataset, but includes all PII.

3. Since UCHC was named as P20WIN’s FERPA Authorized Representative, the plan was to use the Pi Desktop Utility with actual P20WIN data. After P20WIN PLT withdrew this, UCHC had to build a test environment with test data that replicated actual P20WIN data. Each Agency provided UCHC with a “petri-dish” example of their actual data so that test data could be generated with characteristics close to the actual data.

It was with this generated test data that UCHC and OpenClinica tested different configurations of Pi using the Pi Desktop Utility in an iterative fashion until a good configuration of Pi parameters for use in the P20WIN environment emerged. In addition, the P20WIN PLT representative, Jan Kiehne, also ran tests on a sample of the actual data to hone in on the best Pi configuration. After two months, two configurations were presented to the PLT for review and approval. The P20WIN PLT approved Jan Kiehne’s Pi configuration as producing good results for matching in the P20WIN environment. Both Kiehne’s and UCHC’s testing results and proposed configuration are posted for public review.

**End-to-End** - The PLT did not specify any end-to-end test scenarios, so UCHC developed 4 tests to validate the application’s performance in a way that was above and beyond that required by the PLT. A complete query of each data source (four queries) was executed along with a query that mimicked the result needed for the P20WIN Pilot questions. Input data, output data and documentation of validation check for these queries are posted for public review. All five queries were executed correctly and the complete and correct data was built by the query as expected.

**UCHC delivered all test reports to the P20WIN PLT during the final implementation phase of the project, March 2014.**