Archemy™ Showcase Solution

Biobanking Semantics Management Platform

A Big Data and Ontology-Driven Solution

Business Context

The IT group at a major regional Medical Center provides software solutions and services to several research communities, including a cluster of groups that operate collaboratively. The research cluster provides services that facilitate basic, clinical, and translational research and are looking for technologies that facilitate collaboration among their teams.

Initial State

The research community in general places a high value on cross-institutional collaboration. However, existing data, including both research data and records and physical specimens, exists in various formats and is tagged with non-standard terminology and codes, preventing easy sharing with external parties. With a new standardized biobanking data warehouse, records could be made searchable in order to enable and facilitate collaboration with researchers from other institutions.

Archemy™ Solution

Data is sent, either in files or streamed, from the Center's existing data collection systems to a transformation process, which is the focus of the effort. The solution determines the appropriate codes from the designated terminology standard that correspond to values in the internal data. Once this mapping is accomplished, data is sent to the data warehouse system, where it can be queried by external collaborators.

The solution consists of an ETL process followed by an automated transformation process in which metadata, comprised of standard terminology taken from a thesaurus maintained through an ontology manager, is appended to the data.

Technology Employed

Tech Types:	HealthTech - Enterprise Analytics, Enterprise & Sol. Arch., Hyperscale Computing
Ontology Standard:	NCI Thesaurus (an open-source Common Biorepository Model (CBM))
Ontology Manager:	LexEVS (created and maintained by the National Cancer Institute, intended primarily to support NCI Thesaurus and Meta-thesaurus. LexEVS supports W3C Web Ontology Language (OWL))
Translation Service:	Custom SQL procedures (template provided with solution)
Data Warehouse:	Spark SQL
ETL Tool:	Pentaho Data Integration

End State

The solution demonstrates the viability of incorporating the Ontology and automated annotation into a batch ETL process to transform large amounts of data that could ultimately reside in a Hadoop repository.

Reusable Components

- The solution architecture, including integration of the ETL and translation processes