Case Study

Big Data Analytics For Population Health

With the Linguamatics NLP Data Factory, healthcare payers can extract member insights from unstructured big data to improve population risk stratification

QUICK FACTS

Situation: A top-5 payer needed to mine member-related data from a mixture of unstructured formats held in a data lake, to strengthen their analysis of Congestive Heart Failure (CHF) populations. The payer wanted to integrate the extracted data with conventional data warehousing and analytics approaches, to support improved patient stratification.

Solution: The payer implemented the Linguamatics NLP Data Factory with an automation workflow to ingest data from Hadoop, process it and load it into a data warehouse for analysis.

Success: The team demonstrated that the Linguamatics NLP platform can be integrated into existing Hadoop and Netezza systems, to gather and use insights from unstructured data as part of risk stratification analytics for CHF. The NLP Data Factory can be easily extended to support other diseases areas and risk factors such as diabetes, obesity and more.

Situation

Many payers are assessing how to improve stratification of patient populations using big data to fuel the drive toward better member wellness. Risk stratification has so far been biased toward structured data, with major investments in data warehouses, analytical tools, dashboards and Master Data Management (MDM). However, because of the growing availability of electronic health record (EHR) data in Continuity of Care Document (CCD) format from their providers, extensive notes about members and nurses' notes, there is a huge untapped potential in unstructured data. To manage these documents, many groups are making use of Hadoop, as these technologies have proven to scale to the data volumes payers need to support.

But how can payers make effective use of unstructured data to stratify populations more effectively, when much of their infrastructure is tied to structured data, and while the sources of unstructured data are so varied? How can these data worlds be brought together?

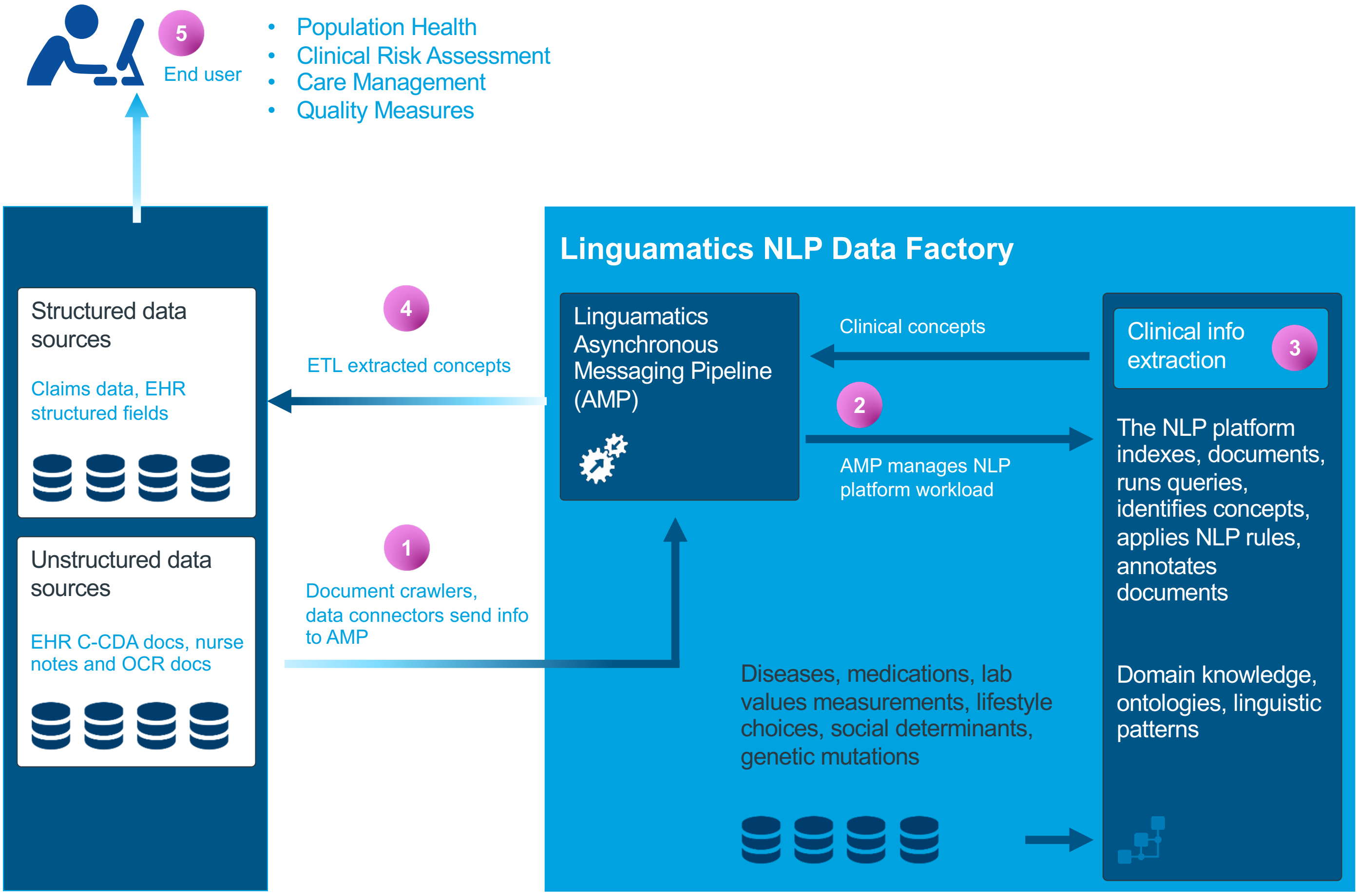
As interest in long-term member wellness increases in importance, it is the insights trapped in unstructured data that will become the differentiator in a changing and competitive market.

Solution

Linguamatics teamed up with a top-5 payer to transform their unstructured data into fuel to drive risk stratification. Unstructured data stored in Cloudera needed to be loaded into Netezza in structured form from CCD, nurses' notes and Optical Character Recognition (OCR) documents.

Unstructured data is extracted from Cloudera Hadoop Distributed File System (HDFS) and passed through an NLP Data Factory to mine risk factors of interest (such as diseases and family history, and lifestyle factors such as smoking), and then turned into structured data. This process is described in more detail in Figure 1.

Figure 1: Technical Workflow


5
End user
Population Health
Clinical Risk Assessment
Care Management
Quality Measures
Structured data sources
Claims data, EHR structured fields
Unstructured data sources
EHR C-CDA docs, nurse notes and OCR docs
4
ETL extracted concepts
1
Document crawlers, data connectors send info to AMP
Linguamatics NLP Data Factory
Linguamatics Asynchronous Messaging Pipeline (AMP)
Clinical concepts
2
AMP manages NLP platform workload
Clinical info extraction
3
The NLP platform indexes, documents, runs queries, identifies concepts, applies NLP rules, annotates documents
Diseases, medications, lab values measurements, lifestyle choices, social determinants, genetic mutations
Domain knowledge, ontologies, linguistic patterns

1 Unstructured member-related documents are sent to the Linguamatics Data Factory via RESTful Web Services to manage information extraction.

2 AMP distributes the documents across multiple Linguamatics NLP servers depending on the required workload. If servers are down, or there are connections issues, AMP will reschedule the extraction jobs.

3 Multiple instances of the platform receive documents; these are indexed and information is extracted. Extracted information may include diseases, medications and lab values, as well as concepts such as Social Determinants of Health, lifestyle factors (smoking, and alcohol and drug use), ambulatory status and living location.

4 AMP sends the extracted data to data warehousing or MDM solutions in XML, JSON or CSV/TSV format.

5 The end user/automated routine is able to run analytics across structured and unstructured data sets to support different business lines.

The NLP Data Factory is used to extract, for example, a person's smoking status to enable them to be grouped by smoking behavior. The different ways this can be represented linguistically are incorporated into the query, and returns a consistent and normalized value associated with each person's status. Standard Extract, Transform, Load (ETL) approaches are used to load the structured data output from the NLP platform into Netezza.

Success

TRANSFORMING UNSTRUCTURED DATA INTO ACTIONABLE INSIGHTS

The team demonstrated that the Linguamatics NLP Data Factory can be integrated into existing Hadoop and Netezza systems to gain insights from unstructured data to be used as part of risk stratification analytics for CHF. Linguamatics helps the payer advance its ability to stratify patients at a much more detailed level of insights. In addition, by cross-referencing insights across multiple sources of unstructured data, a more complete picture emerges.

LEVERAGING EXISTING INFRASTRUCTURE

By providing an extraction pipeline that supports existing investments in big data and data warehouses, rather than tearing out well understood and established approaches, the Data Factory is able to plug natural language processing into these systems, to enhance understanding of members based on unstructured data.

FAST TIME TO VALUE

The payer learned how to build queries very quickly due to the platform's GUI-driven NLP interface—users do not have to be NLP experts. The system is fully configurable, so that modifications can be easily made without Linguamatics Professional Services.

EXTENDING VALUE INTO OTHER DISEASE AREAS AND APPLICATIONS

The platform infrastructure can be easily extended to support other disease areas and risk factors—for example, COPD, diabetes and obesity. Investigations into reducing the manual chart review required for HEDIS (Healthcare Effectiveness Data and Information Set) metrics, and improving the capture rate, are also being addressed with Linguamatics NLP, demonstrating the power and flexibility of this enterprise platform.

CONTACT US

+44 (0)1223 651 910 (U.K.) | +1 617 674 3256 (U.S.)
nlp@iqvia.com
linguamatics.com