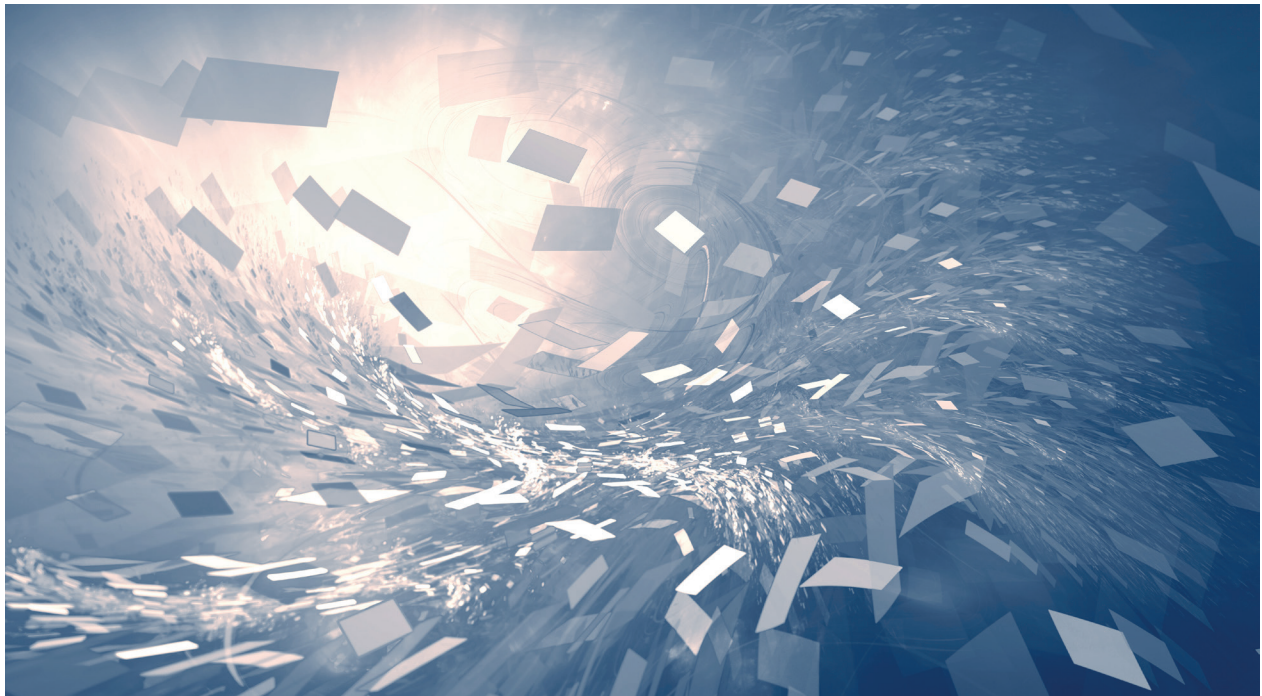


# Linguamatics I2E AMP for Extract, Transform, and Load: Incorporate key concepts from unstructured data



## Challenge

Integrating data from a variety of sources into a data warehouse or other data repository centralizes business-critical data, and speeds up finding and analyzing important data.

The Extract, Transform, and Load (ETL) process of extracting data from source systems and bringing it into a data warehouse is well established. While many ETL tools can handle structured data, very few can reliably process unstructured data and documents.

## Solution

Linguamatics I2E AMP is an ideal tool for ETL processes from unstructured or semi-structured data formats, particularly with high, variable, and unpredictable throughput, and where different input document types (e.g. email, PDF, and text) require different workflows.

I2E AMP runs queries that reliably and repeatably extract specific data points from a particular unstructured document type. I2E AMP supports business decisions about which data items are required, and what extracted data is loaded into the database or master data management system.

## Benefits

Using I2E AMP for ETL, enterprises can:

- ◆ easily add insights from unstructured data into data warehousing environments;
- ◆ enhance existing investments in warehouses, analytics, and dashboards;
- ◆ quickly add and modify processes; and
- ◆ recover from server crashes without losing data.

## Key features

I2E AMP provides precise, comprehensive, and reliable data transformation, identifying key concepts in unstructured text, and normalizing the terms found to standard, primary labels. This more standardized semantic data representation enables integration of extracted concepts with data from structured sources.

Using I2E AMP, organizations can:

- ◆ support a traditional SQL database, a triple store, a NoSQL database, an enterprise-class data lake

such as Cloudera or MarkLogic, or a machine learning platform;

- ◆ provide high-throughput, secure, fault-tolerant, scalable, real-time I2E document processing ETL for data warehouses;
- ◆ manage multiple I2E servers for annotation and querying, by load-balancing resources and buffering high, variable, and unpredictable document input streams; and
- ◆ integrate I2E into a service-oriented architecture (SOA) via a RESTful API.

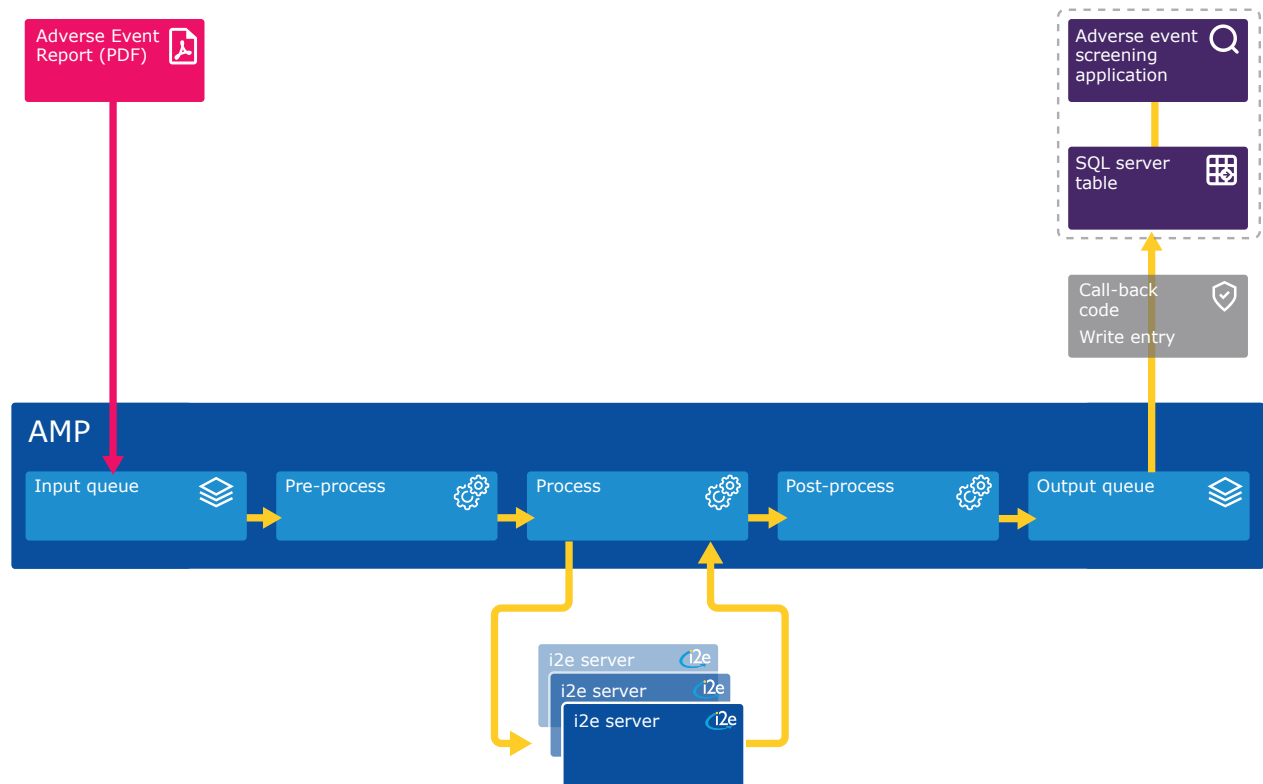


Figure 1: This example I2E AMP for ETL application is populating adverse event databases from unstructured documents. I2E AMP ingests the adverse event and case reports, identifies and extracts the required data, transforming it into a standard or normalized form, and loads it into the database for review and analysis.

Other applications of I2E AMP with ETL include: extracting lifestyle choices such as smoking status, or clinical insights such as ejection fraction, from patient notes, to feed clinical risk applications; and extracting the most recent gene-disease-biomarker associations from literature and patents, to feed a translational discovery database.

If you are interested in learning more about I2E AMP and ETL, please contact us at [enquiries@linguamatics.com](mailto:enquiries@linguamatics.com)