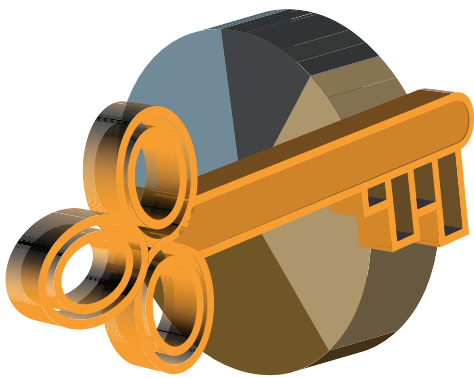# PolyAnalyst™ 6.5

TECHNICAL CAPABILITIES AND SYSTEM REQUIREMENTS

PolyAnalyst™ is comprehensive data and text mining system providing tools for handling all steps of a typical data analysis process: from data loading, integration, manipulation and cleansing, to advanced text analytics, machine learning and knowledge discovery, and to versatile results visualization and reporting. Tools for the analysis of structured and textual data in a single environment enables joint processing of all available data for better insights.
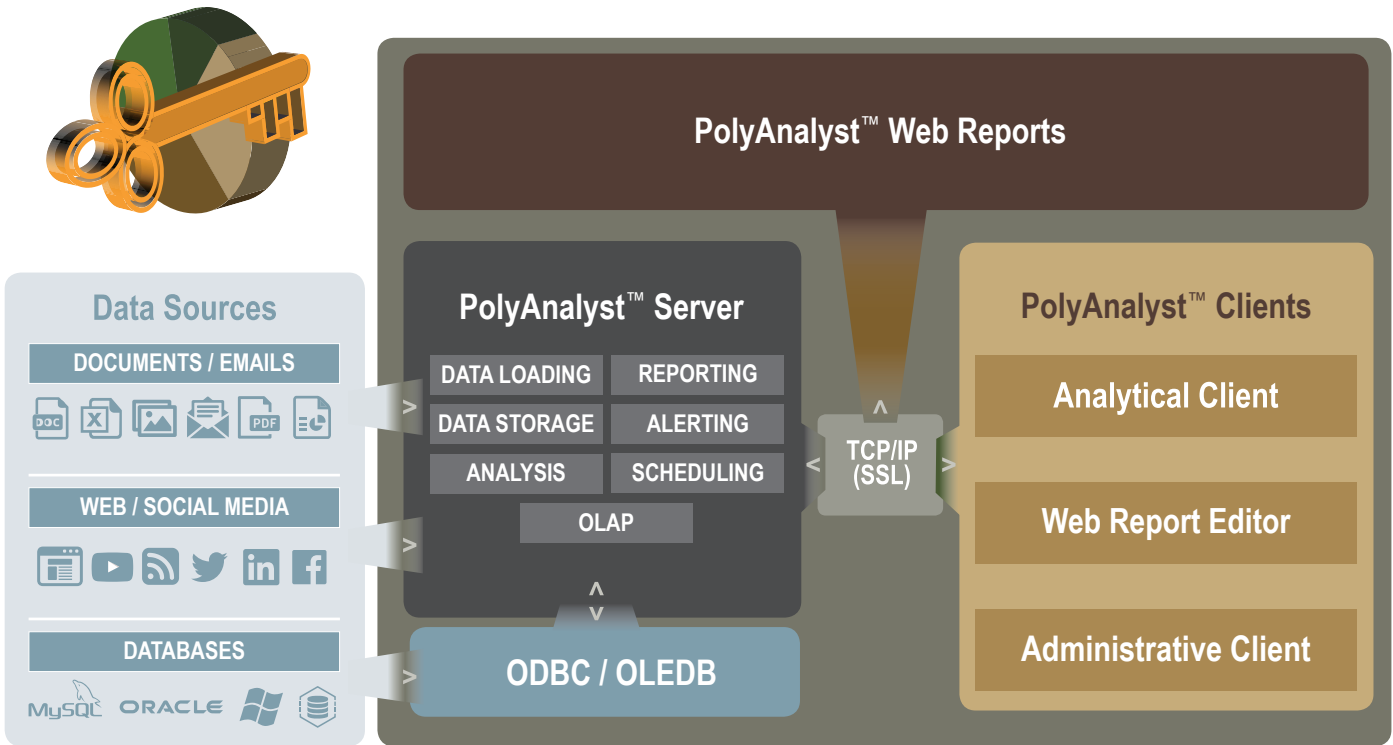
Figure 1. PolyAnalyst architecture.

Client/Server architecture

Server architecture enables the implementation of PolyAnalyst™ 6.5 as an enterprise level analytical system. It facilitates the collaboration between data analysts working on the same projects and sharing various related resources such as analysis scenarios, dictionaries, taxonomies, and multi-dimensional matrices. Server architecture helps enhance the performance of the system by performing calculations on the most powerful machines, reducing data transfer over the network, scheduling execution of tasks at a given time, and generating custom reports and condition based alerts for different groups of business users. It provides the centralized management and audit of the list of system users and their actions.

Security

Recognizing that data is one of the most valuable and sensitive assets of a modern organization, PolyAnalyst™ provides solid mechanisms to ensure data security. Communications between client and server are performed in a fully encrypted manner with a new encryption key generated by the server for every communication session. PolyAnalyst supports secure user login based on user rights and passwords and keeps track of individual and group rights and sequences of actions carried out by the users. In addition, full compliance with the requirements of the HIPAA legislation facilitates PolyAnalyst implementations at healthcare and insurance organizations.

Scalability and Performance

PolyAnalyst provides industrial level scalability: it can handle huge amounts of data within reasonable time intervals. This scalability is ensured through a combination of several factors. PolyAnalyst™ utilizes hard disk instead or RAM for holding all data and supporting meta-information. Special scalable implementation of analytical algorithms enables the system to process large volumes of data. PolyAnalyst user interface allows the user to develop complex data analysis scenarios without loading data in the system, thus saving analyst's time. The availability of PolyAnalyst as the first 64-bit analytical system and its server farm implementation help dramatically increase the performance of the system and allow numerous users to perform data analysis and report viewing simultaneously.

Reusable analysis scenarios

A typical data analysis project involves a sequence of steps of data loading, preparation, analysis and reporting. Frequently, the same sequence of analytical steps has to be carried out repetitively on new batches of data. PolyAnalyst™ is a self documenting system that provides intuitive visual tools for developing and editing reusable multi-step data analysis scenarios. The system is easy to learn and fun to use. Data analysis and report generation scenarios can be scheduled for re-execution at any given time. This ensures that business users have timely access to up-to-date reports built on most recent data.

Data loading and integration

Whatever your data sources are, PolyAnalyst™ provides means for loading and integrating these data. PolyAnalyst™ can load data from disparate data sources including all popular database, statistical, and spreadsheet systems. In addition, it can load collections of documents in html, doc, pdf and txt formats, as well as load data from an internet source including websites, RSS feeds and Social Media. PolyAnalyst™ offers visual on-the-fly integration and merging of data coming from disparate sources to create data marts for further analysis. It supports incremental data appending and referencing data sets in previously created PolyAnalyst™ projects.

Data manipulation

Every data analysis project starts with data normalization, manipulation, and exploratory analysis. In fact, the dominant part of the analyst's time is spent on data manipulations preceding the application of machine learning algorithms. PolyAnalyst™ provides the analyst with a vast set of powerful manipulation tools for data cleansing, aggregation and derivation of new attributes. Virtually any data transformation task can be solved through sequential application of PolyAnalyst™ data manipulation tools.

Analytical algorithms

The main quality the user expects in a knowledge discovery system is its ability to use mathematical algorithms to learn from historical data and predict outcomes of future situations. PolyAnalyst™ provides a broad selection of analytical algorithms for clustering and categorization of data, solving prediction, link analysis and affinity grouping tasks, learning patterns and discovering anomalies in data. Ranging from neural networks and decision trees, to Bayesian Networks and Support Vector Machines, to CHAID and logistic regression, and to Case-based Reasoning and Evolutionary Programming, PolyAnalyst™ scalable algorithms enable the user to solve the analytical task at hand.

Text Analysis

The data might contain attributes holding free form text, like in incident reports, claims notes, or in survey responses. Or the project might require the analysis of huge collections of documents in various formats, possibly harvested from the Internet in real time. Whatever the task, PolyAnalyst™ offers a collection of text analysis algorithms that enable the data analyst to solve it. Based on Megaputer linguistic platform and the incorporation of various semantic dictionaries, PolyAnalyst™ represents a powerful natural language processing tool. PolyAnalyst™ provides tools for handling both analyst driven analysis, such as taxonomy based categorization and Text OLAP, and data driven analysis, such as intelligent spell checking, keyword, and entity extraction, clustering, and taxonomy creation.

PolyAnalyst 6·5 Features

LOADING	CLEANSING + MANIPULATION	TEXT ANALYSIS + MACHINE LEARNING		REPORTING
Databases	Pattern Substitution	Document Categorization	Link Analysis	Report Templates
Spreadsheets	Intelligent Spellchecker	Taxonomy Generation	Anomaly Detection	GIS
Statistical Systems	Dictionary Management	Entity Extraction	OLAP	Report Publishing
Document Collections	Grammar Checker	Text OLAP	Trend Analysis	Web Reports
Emails Flat Files	Data Sources Integration	Semantic Search	Categorization	Scheduled Execution
Internet	Aggregation	Document Clustering	Affinity Grouping	Executive Reporting
Hadoop	Attribute Derivation	Keyword Extraction	Clustering	Trends & Alerts
Social Media	Anonymize	Pattern Detection	Prediction	Drill-Down Reporting

Figure 2. PolyAnalyst™ key capabilites.

Multi-dimensional analysis

The development of multi-dimensional cubes based on the analyzed data allows the user to answer a variety of business questions by slicing data across various dimensions. This technology is widely known as OLAP (for on-line analytical processing). PolyAnalyst™ offers a robust OLAP engine as one of its data analysis algorithm. In addition, PolyAnalyst™ enhances standard OLAP by adding dimensions defined on fields containing free form text and offering a unique multi-dimensional OLAP interface.
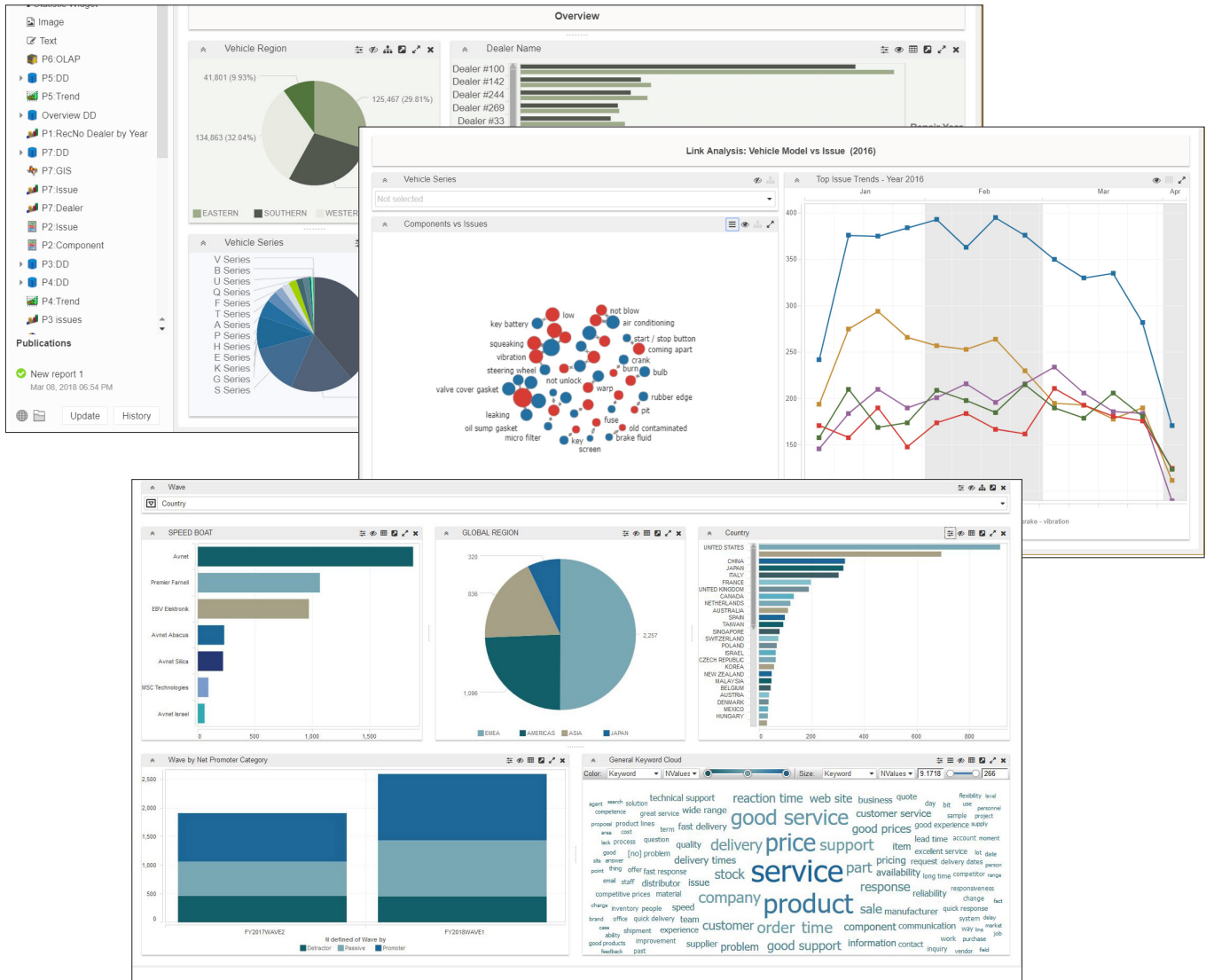
Figure 3. *PolyAnalyst™ web reports.*

Interactive visualization

PolyAnalyst provides the data analyst with an immediate feedback on the results of their analysis. It offers interactive and visual user experience whenever possible. In fact, exploratory analysis represents a very important stage of the complete data analysis cycle. The user of PolyAnalyst™ commands a variety of tools for interactive visualization of the data and the results of the analysis: one and two-dimensional histograms, pie charts, line charts, bubble charts, scatter plots, snake charts, statistical widgets, link charts, and trends graphs help the user make sense of the data.

Reporting

PolyAnalyst enables the data analyst to create custom reports delivering key results of the analysis to business users across the organization in a clean, consistent and easy to comprehend format. Interactive reports include a mixture of graphs, tables, numbers, text and links to other PolyAnalyst objects. Reports can be scheduled for re-execution at a given time to provide business users with results based on the analysis of the most up-to-date data. Static snapshots of reports can be exported to PDF, HTML and RTF format.

1. Data loading and integration

    a. Data Sources: ODBC, OLEDB, XML, CSV, MS Excel, Web, File System, FTP

    b. Document Formats: PDF, ASCII, HTML, MS Word, MS RTF, RSS Feeds

    c. Character Formats: ASCII, Latin-1, Double-byte, UTF

    d. Web Data sources

    e. Direct Email Import (PST or NSF)

    f. Email Server Data Source

    g. Social Media sources:
        i. Twitter and Facebook APIs
        ii. Webhose, BrandWatch and SM2 APIs

    h. Dataset Links
        i. Federated Search

2. Data integration

    a. Data joins on sets of keys

    b. Data merging

    c. Referencing data sets in other PolyAnalyst™ projects

    d. Exporting results of the analysis to external RDBMS

3. Data cleansing, manipulation and exploratory analysis

    a. Attribute name, type and value mapping

    b. Data aggregation

    c. Data consolidation

    d. Expanding and collapsing transactions

    e. Data transformations

    f. Sampling and partitioning

    g. Derivation of new attributes

    h. OLAP

    i. Statistics
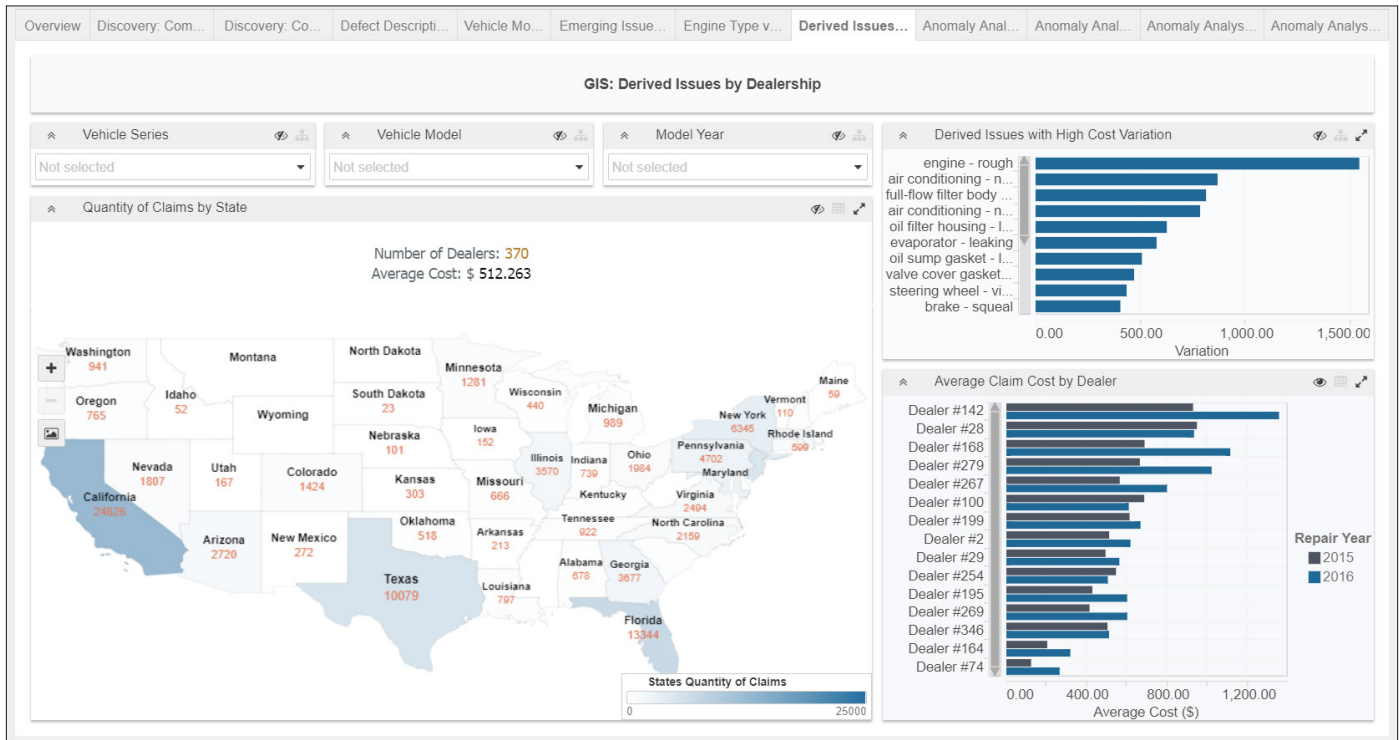
    j. Link analysis

    k. Geospatial mapping
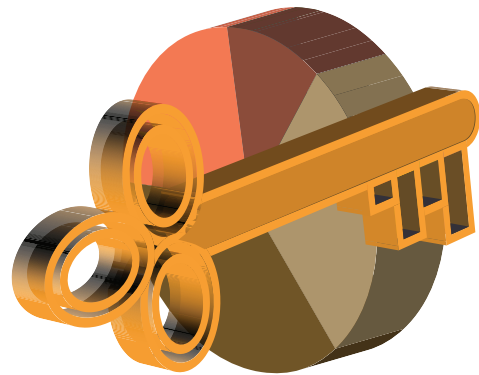
Figure 4. PolyAnalyst geospatial mapping – state level.

4. Predictive modeling and machine learning

a. Classification
   i. Decision Tree
   ii. Naïve Bayesian
   iii. Boosting
   iv. Bayesian Network
   v. R-Forest
   vi. CHAID
   vii. Case-based Reasoning
   viii. Neural Network
   ix. Support Vector Machine
   x. Instance-based Learning
   xi. Logistic Regression
   xii. MARS (Multivariate Adaptive Regression Splines)

b. Clustering
   i. CHAID
   ii. Localization of Anomalies
   iii. Kohonen Self-Organizing Maps

c. Prediction
   i. Neural Network
   ii. Evolutionary Programming
   iii. Linear Regression

d. Anomaly detection
   i. Distribution Analysis
   ii. MediCop
   iii. Data Audit

e. Affinity grouping
   i. Market Basket Analysis
   ii. Visual Link Analysis
   iii. Discriminant Basket Analysis

f. Dimension Reduction
   i. Factor Analysis
   ii. Principal Component Analysis
   iii. Entity Resolution (Fuzzy Matching)
   iv. Correlation Analysis

g. Time Series Analysis
   i. Additive Time Series Analysis

h. Network Analysis
   i. Social Network Analysis
   ii. Link Analysis

5. Text Mining

a. Text cleansing and normalization
   i. Near duplicates detection
   ii. Fragment Analysis
   iii. Editable dictionaries of phrases, stop-words, synonyms and other semantic relations
   iv. Intelligent spell checking and correction
b. User driven analysis
   i. Information retrieval with search queries of any complexity
   ii. Query expansion with semantic relations (synonyms, hypernyms/hyponyms, etc.)
   iii. Taxonomy-based categorization
   iv. Automated categorization based on pre-set examples
   v. Standard entity extraction
   vi. Custom entity extraction
   vii. Relationship extraction
   viii. Finding similar documents
   ix. Multi-dimensional reporting from text data (Text OLAP)
c. Data driven analysis
   i. Language detection
   ii. Part of speech tagging
   iii. Keyword extraction
   iv. Phrase extraction
   v. Syntactic analysis: Dependency parser
   vi. Syntactic analysis: Constituency parser
   vii. Standard entity extraction
   viii. Text clustering (binary and hierarchical)
   ix. Taxonomy generation from the analysis of raw data
   x. Sentiment analysis
   xi. Visual cluster analysis
   xii. Classification model training based on pre-categorized examples
d. Languages
   i. Support for 16 additional languages: *European:* English, Spanish, French, Russian, German, Dutch, Portuguese, Italian, Greek, Turkish, and Polish *Asian:* Chinese, Japanese, Korean, Thai, and Arabic
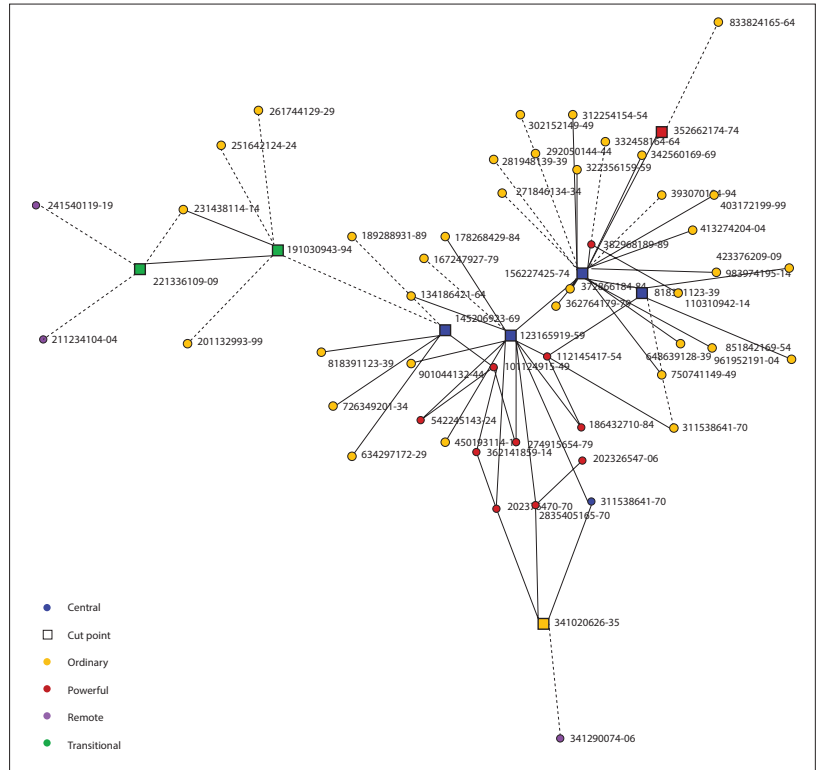

Figure 5. Visualization of the Social Network Analysis results.

e. Extras
   i. Negation detection
   ii. Anaphora resolution
   iii. Modality detection
   iv. Abbreviation expansion
   v. Pattern Definition Language
   vi. Regular expressions
f. Customization: support for domain-specific dictionaries
   i. RxNorm
   ii. MeSH
   iii. SNOMED CT
   iv. MedDRA
   v. Eagle-i
g. Customization: domain-specific taxonomies for
   i. Insurance
   ii. Healthcare
   iii. Hospitality
   iv. Fast Food
   v. Manufacturing

6. Graphical presentation of results with interactive drill-down feature

- a. Histograms (2D and 3D)
- b. Scatter plots
- c. Line charts
- d. Bubble charts
- e. Thermal charts
- f. Rule graphs
- g. Link charts
- h. Snake charts
- i. Trends graphs

7. Reporting

- a. Custom reporting templates
- b. Reports summarizing up-to-date results
- c. Interactive drill-downs for business users
- d. Browser-based reports
- e. Parameterized reports
- f. Persistent report instances
- g. Publishing reports to popular document formats

8. Analysis scenarios

- a. Visual development of reusable analysis scenarios
- b. Drag-and-drop development of reports
- c. Group nodes feature for managing large scenarios

9. Scheduling

- a. Automated execution of analytical scripts at given times
- b. Condition-based alerts
- c. Generation of informative email notifications
- d. Automated publishing of reports

10. Scalability

- a. Client/Server implementation
- b. Client – Server communications over TCP/IP protocol
- c. Utilization of hard disk instead of RAM
- d. Scalable implementation of algorithms
- e. 64-bit implementation available
- f. Analytic scenario development prior to actual data loading

11. PolyAnalyst Hadoop integration

- a. Working with data in HDFS
- b. Data does not enter PolyAnalyst
- c. Analysis is performed on Hadoop cluster
- d. PolyAnalyst can export data to Hadoop
- e. Implemented text indexing, taxonomy categorization, machine learning, and more
- f. Blazing speed analysis

12. Usability

- a. Interactive drag-and-drop experience throughout the system
- b. Visual development of reusable data analysis scenarios
- c. Tight integration of analytical and reporting applications
- d. Visual creation of nice looking interactive reports for business users
- e. Publishing reports to popular document formats for better collaboration
- f. Group nodes
- g. Subject areas

13. Security

- a. User name and password based authentication
- b. Support for LDAP and MS Windows based authentication
- c. Fully encrypted client-server communications
- d. User activities logging
- e. Compliance with HIPAA regulations

PolyAnalyst™ Server

Minimum Server Requirements:

1.4 GHz or higher: dual-core Intel or AMD CPU
8 GB RAM
10 GB of free hard drive space
Windows 7, 8, 10, 2003, 2008, 2012, or higher

Recommended Server Resources:

3 GHz or higher: six core+ (Intel or AMD)
32 GB or more, additional RAM may be needed
    depending on the number of expected concurrent
    users and complexity of projects
1 TB RAID, *SSD is recommended*
Windows 8, 10, 2008, 2012 or higher

PolyAnalyst™ Clients

Minimum Client Requirements:

1 GHz or higher: dual-core+ Intel or AMD processor
2 GB or more RAM
1 GB or more of free hard drive space
Windows 10, 8, 2008, 2012 or higher
Browser: Chrome (2016 or newer), Firefox (2016
or newer), or MS Edge

Recommended Client Resources:

2 GHz or higher: dual-core+ Intel or AMD processor
4 GB or more RAM
2 GB or more of free hard drive space
Windows 10, 8, 2008, 2012 or higher
Brower: Chrome (latest version), Firefox (latest version),
or MS Edge (latest version)

PLEASE NOTE: System specifications may vary based on project variables, i.e. size of team, the amount of data, number of projects and more. These specifications should be used as a starting point for initial implementation of PolyAnalyst. Resources may need to be audited when the project volume expands and/or the amount of end users increase.

Megaputer

Your Knowledge Partner™

info@megaputer.com
+1 812-330-0110

www.megaputer.com