psydex

What The World Is Thinking®

TIME CRITICAL INSIGHT *IN A REAL TIME WORLD*

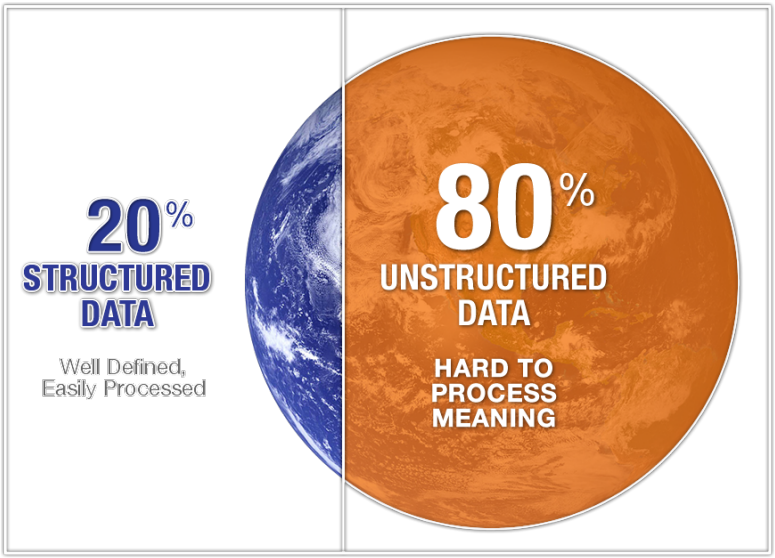
THE PROBLEM

The world will produce a ZETABYTE (1 billion terabytes) of new data this year.

20% is easily processed while the other **80%** is news, social media, emails, messages and other forms of ***un-structured*** data...

...and it is all happening in ***real time***.

Databases and batch processing systems like Hadoop were not designed to analyze ***un-structured*** data in ***real time*** for ***patterns***, ***trends*** and ***signals.***



For decades, companies have operated under a model of collecting data and putting it to rest and optimizing it for structured query. This is appropriate for well-structured, transaction oriented systems, but for systems that process high volume, high variety and high velocity data this does not work if the data needs to be analyzed in a time critical fashion. Relational databases have never been intended to store, index and analyze un-structured data, and SQL was certainly not designed as a query language for un-structured data. Un-structured data is typically void of a data model and semantic tokens, and related information cannot be effectively indexed to accommodate ad hoc queries and analytics. Ad hoc semantic queries against relational data are slow and virtually impossible against large historical data sets requiring retrospective analytics, time series, pattern and trend generation.

Regardless of the data type, databases have generally proven ineffective at performing intensive analytic operations under heavy ingest/query loads. Columnar data stores are an incremental improvement to relational databases, and solid state drives have reduced seek times and increased transfer rates on disk drives, but disk I/O remains orders of magnitude slower than memory based operations.

New batch processing approaches, such as Hadoop, are suitable for batch analytics against large, relatively ***static*** data sets, but are not designed to meet the demands of analyzing and processing streaming data in a time critical fashion. Reengineering Hadoop to accomplish this will prove to be very challenging and require years of development to be production ready for mission critical systems.

A problem with most information systems that handle un-structured data is the tight coupling of taxonomies with data storage and indexing. In an age where social media language and lexicons are constantly evolving, systems that tag and store semantic information at the point of ingest/indexing cannot easily adapt to changes in the way people communicate or handle newly discovered knowledge. In these systems, new language or vocabulary cannot be efficiently searched and analyzed against *historical* data without re-indexing or reloading. This is a significant problem; for example, when an analyst discovers previously unknown jargon that means something, the analyst may intuitively want to back-test this to validate a hypothesis.

3343 Peachtree Road, N.E., Atlanta, Georgia 30326 | www.psydex.com | info@psydex.com | 404-551-5238

A NEW APPROACH - NEXT GENERATION

We all know disk I/O is slow relative to RAM and the cost disparity between disk drives and memory capacity continues to decrease. So why build and store indexes of un-structured data in a database or on a disk based file system? The data needs to be put to rest, but could indexes reside All-In-Memory™ (AIM™)?

What if AIM™ Indexes could be organized in a way to accommodate any imaginable semantic query based on relative or temporal proximity of tokens. Using a token level indexing approach, a system could be language-independent.

There are of course limits to physical memory and process spaces running on a single computer, but what if an index could be partitioned across many nodes? What if analytic services operate against All-In-Memory™ indexes and never have to read or write to disk? What if services could be broken up into smaller tasks, all managed to achieve maximum scale and performance across thousands of commodity computers?

There are huge challenges to an approach like this because many machines would be required to have enough memory available to achieve the scales possible with disk based indexing. There would need to be a way to seamlessly manage physically separate memory across many nodes, and with new data constantly streaming in, this would create unique challenges. Could it be done? Could the collective system memory be managed as a huge FIFO buffer functioning seamlessly across many nodes?

The Psydex AG (Analytics Grid) was designed and architected from the bottom up to address all the problems and challenges discussed above.

PSYDEX AG

Psydex AG (Analytics Grid) is a next generation SOA platform for *real time* search, data mining and predictive analytics across large, fast-moving un-structured data streams. Psydex is based on a novel All-In-Memory™ (AIM)™ approach that leverages many computers operating in parallel.

- RUNS ON COMMODITY HARDWARE
- MASSIVE PARALLEL PROCESSING (**MPP**)
- All-In-Memory™ (**AIM™**) INDEXES
- ***DATA*** | ***INDEXES*** | ***MODELS*** ARE *DE-COUPLED AND CAN CHANGE INDEPENDENTLY*

AG fully supports indexing multi-dimensional structured data, but excels at hard-to-process, fast-moving un-structured data streams where there is little or no structure and most importantly; where the data needs to be analyzed in a ***time critical*** fashion.

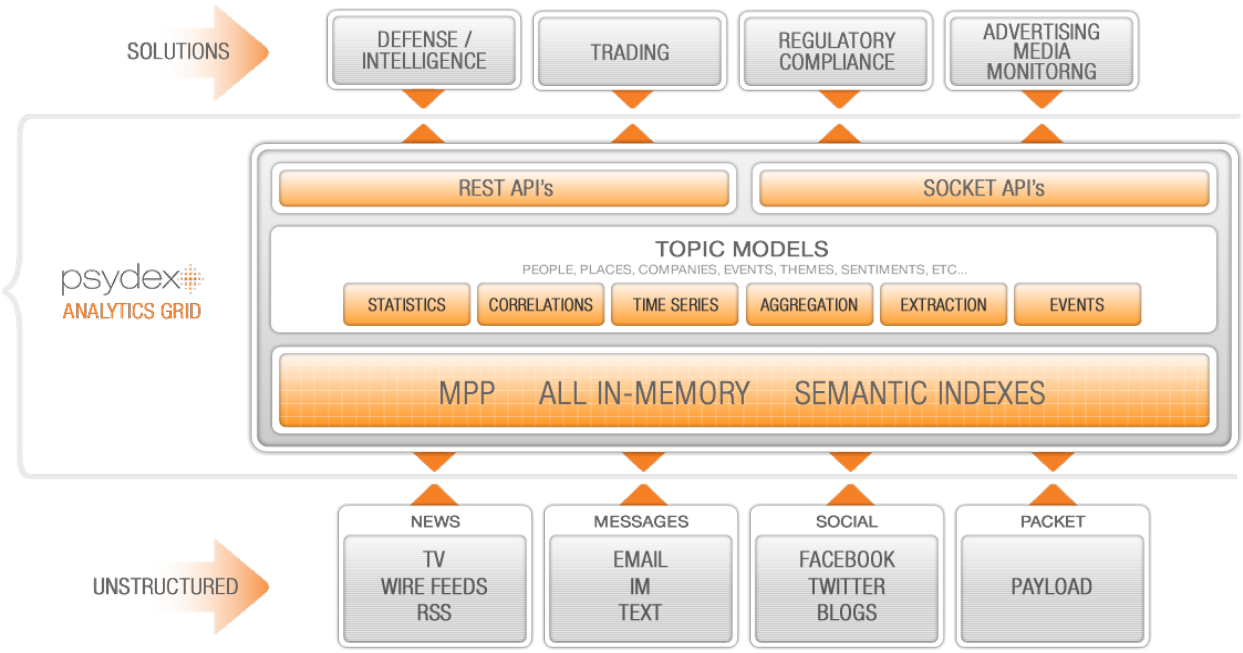
3343 Peachtree Road, N.E., Atlanta, Georgia 30326 | www.psydex.com | info@psydex.com | 404-551-5238

psydex

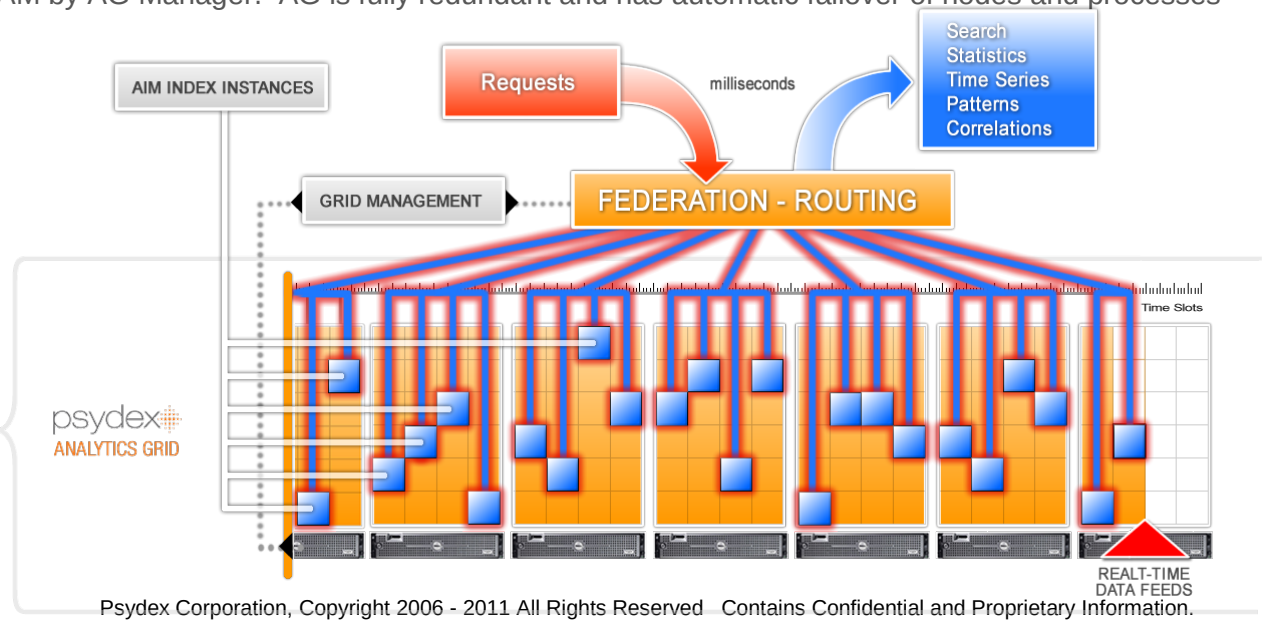
What The World Is Thinking®

TIME CRITICAL INSIGHT *IN A REAL TIME WORLD*

AG can perform complex ad hoc semantic queries against petabytes of data in milliseconds. AG is a powerful complement to existing relational database systems; acting as a completely independent real time Index system. AG is at least one thousand times faster than batch processing systems such as Hadoop, at analyzing real time, streaming unstructured data.



AG is based on a MPP (Massive Parallel Processing) Shared-Nothing Architecture with Indexes and Services partitioned and spread across ***many*** process spaces on ***many*** Nodes. Collectively AG Indexes are managed as a FIFO buffer, ingesting streaming data from queues, databases or directly from native feeds. As Psydex AG memory reaches capacity, older process spaces are swapped out of RAM by AG Manager. AG is fully redundant and has automatic failover of nodes and processes



3343 Peachtree Road, N.E., Atlanta, Georgia 30326 | www.psydex.com | info@psydex.com | 404-551-5238

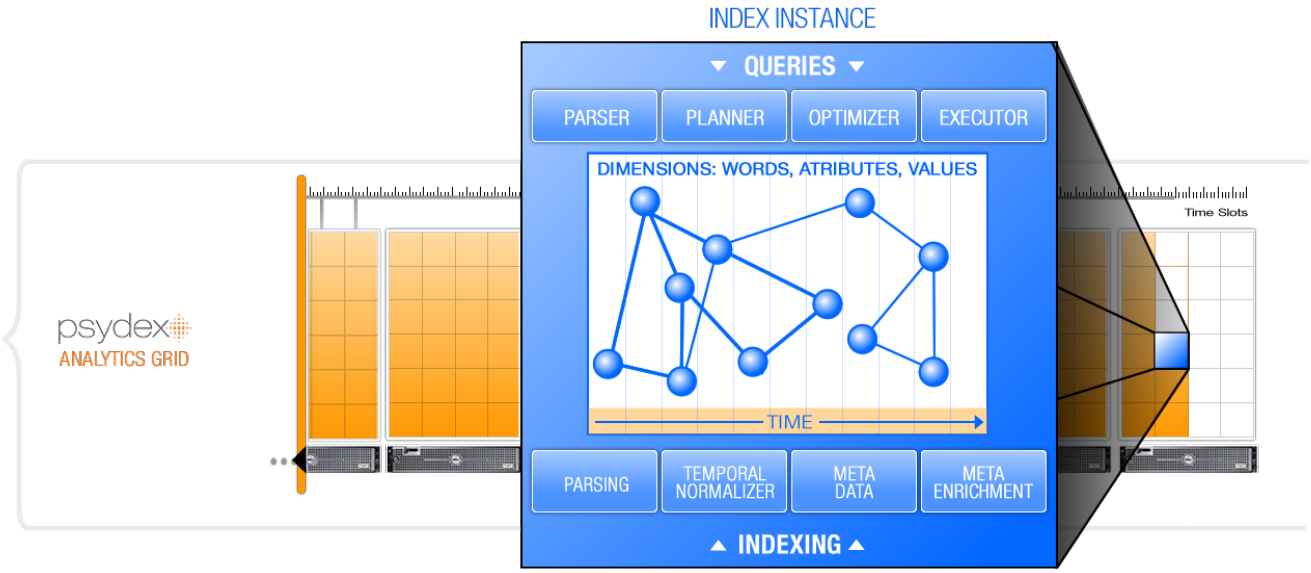
PSYDEX AG - ANALYTICS

AG provides seamless, real time analytics over both streaming and historical data - bridging the gap between complex event processing and data warehousing/business intelligence systems. Signals and alerts are generated in real time from time series, statistics and correlations using Topics Models based on PQL expressions (see Psydex Query Language). The AG Event Processor analyzes inputs such as Topic Model counts, statistics, correlations, co-occurring Topic Models and time series to identify higher-order events with extreme speed and accuracy.

All-In-Memory™ (AIM™) Indexes

Indexes in AG are represented as All-In-Memory™ multi-dimensional graphs of tokens (e.g. words, attributes, and values) organized around time and partitioned by time & source across the grid. Indexes handle source-specific metadata, transformation and enrichments upon loading and are dynamically deployed, managed and executed in AG as All-In-Memory ™ instances. Each source in AG is partitioned into many Index Instances, each representing a time slice of a source. A typical source Index might have hundreds of instances functioning seamlessly across many nodes. Each instance is multi-threaded and typically resides in a single 4+ GB process space, depending on the operating system. Operational parameters are configurable at run-time.

Federators and the AG Manager communicate with Index Instances to perform intelligence routing of requests. Time Series and aggregate functions occur locally in each process space. All Index instance functions are multi-threaded and wire format for communications is based on a proprietary fast serialization protocol.



3343 Peachtree Road, N.E., Atlanta, Georgia 30326 | www.psydex.com | info@psydex.com | 404-551-5238

TIME CRITICAL INSIGHT *IN A REAL TIME WORLD*

Topic Models

Topic Models describe things (e.g. people, companies, places etc.) and also abstract concepts such as "Fear" and "Civil Unrest". Topic Models are represented by expressions (semantic rules) comprised of characters, words, phrases and operators that define relative or temporal proximity with other words and phrases. Topic Model expressions are defined using the Psydex Query Language or PQL (see PQL) and are always processed in an ad hoc fashion, enabling models to evolve independent of data and indexes.

Topic Models are completely separate from indexes and can be constructed at run-time to evolve as language, taxonomies and knowledge evolves. Versioning and security permissions are supported for Topic Models, and they can be updated and viewed in a historical context without having to re-index or re-organize indexes.

Time Series

Time series represent source-weighted frequency counts for Topic Models across varying period & interval combinations (e.g. 1 day/1minute, 30 day/daily). Time Series counts are based on custom source weightings that can be defined by users. Time Series can be analyzed like stocks using stochastic, Fibonacci and other statistical algorithms.

Statistics

Statistics represent current values (e.g. .mean, standard deviation, Z-Score, exponential moving average) for Topic Models across varying period & interval combinations (e.g. 1 day/1minute, 30 day/daily). Every token that moves through AG causes statistical updates for all matching models. Statistics are based on custom source weightings that can be defined by users.

Correlations

AG cross correlates structured and un-structured data streams in real time and also performs auto/self correlations to discovery patterns within a stream. Correlations can be calculated in-phase or using various phase shift increments. This is useful in determining cause/effect where lags are involved.

Discovery

Discovery services combine Indexes, statistics, correlations and other functions to generate real time insight in to streaming data sources simply by observing the stream and learning from it.

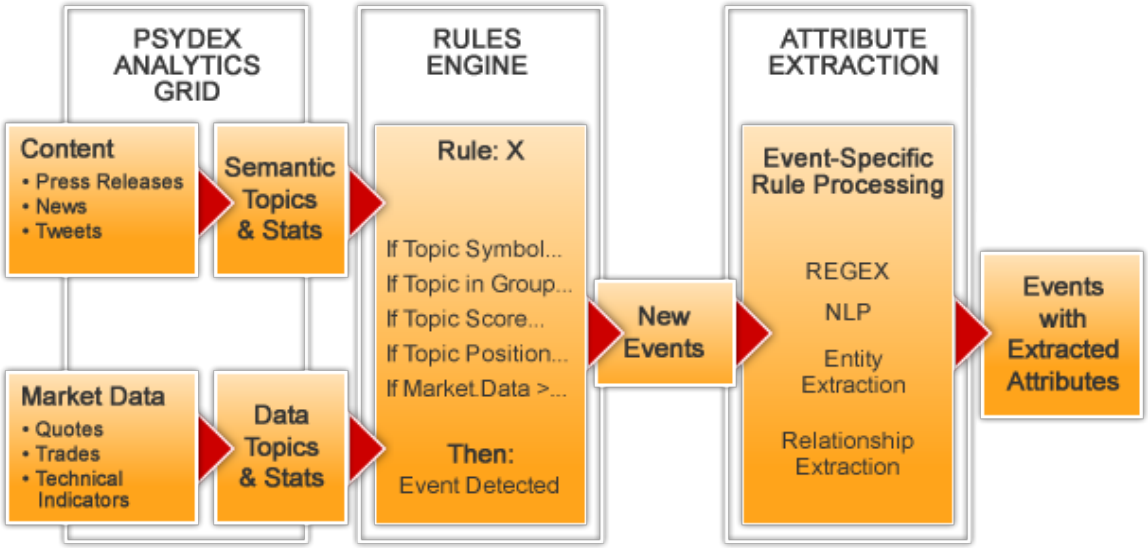
Search

Search enables real time "search" to determine high frequency word/phrase counts mentioned around other words/phrases. Search is similar to Google-like "suggestions" but operates on streaming data. Search also retrieves content based on identifiers, search phrases and data ranges.

3343 Peachtree Road, N.E., Atlanta, Georgia 30326 | www.psydex.com | info@psydex.com | 404-551-5238

PSYDEX AG – EVENT PROCESSOR

AG Event Processor incorporates Topic Model time series, statistics and correlations to identify real-world events with high precision and speed. The detection of unusual patterns in real time is typically the first stage in AG's highly extensible process that employs domain-specific rules and powerful capabilities for extracting facts and figures.



Rules Engine

AG Event Processor leverages commercially available rules engines (e.g. JBOSS Rules/Drools) and also partners with CEP (Complex Event Processing) providers. Rules are organized around domain-specific event types and involve relative and temporal proximity of Topics Models across space and time, statistical thresholds, association with groups, classes and other topics, as well as conditional criteria. Satisfied rules result in the extraction of rule-specific attributes and values that drive downstream processing and decision making.

Attribute Extraction

Whether you need to identify the companies involved in an acquisition, or the location and magnitude of the latest earthquake, AG Event Processor extracts facts and figures with high precision. When an event is detected, a set of rule-specific attribute extraction functions are applied. Attribute extraction leverages best of breed techniques and technologies, including regular expressions, Natural Language Processing (NLP), entity and relationship extraction to identify and extract attributes that define the event. AG Event Processor supports multiple extraction functions for each event attribute to achieve the highest precision based on the source and format of the content.

3343 Peachtree Road, N.E., Atlanta, Georgia 30326 | www.psydex.com | info@psydex.com | 404-551-5238

PSYDEX AG - SOURCE ADAPTERS

Source adapters in AG are easily customized to handle a wide variety of streaming/real time and historical data sources. An SDK exists for writing custom source handlers, parsers and loaders. Source adapters can be easily updated and deployed with no impact to historical or real time data ingest. AG can initialize all associated Index instances for a sources in just minutes by breaking up loading into many independent parallel jobs, and can process sustained, real-time inflows of streaming data by instantiating AIM™ Index Instances in an intelligent and manageable way.

Source adapters exist for JMS, JDBC, Reuters Data Feed (RDF), HTTP/S, NNTP, FTP, ICE Impact Data Feed etc... Parsers are available for many formats and encodings, including HTML, RSS, NewsML, NITF, XML, ICE Impact Data Format, Reuters Open Message Model (OMM), PCAP etc...

Sampling of available Index Adapters (Structured and Un-Structured):

- ### Network Traffic
  TCPDump, WireShark, Bro, Snort, Snoop, PCAP

- ### Market Data
  Thomson Reuters RDF, Nasdaq, NYSE, ICE Impact Data Feed

- ### Enterprise
  Email, File Systems, Document Stores, JDBC

- ### News Wires (Low-Latency Proprietary Feeds)
  Associated Press, Businesswire, DOW Jones, MarketWire, PR NewsWire, Thomson Reuters

- ### TV Captions
  Currently Captioning Live: Bloomberg, CNBC, CNN Headline News, CSPAN, ESPN, ESPN News, ESPN 2, FOX Business Network, FOX News, History Channel, MSNBC, Weather Channel (Other CC Channels area easily configured)

- ### Internet/Social Media/News
  Twitter, Facebook, Yammer, Atom, RSS, Blogs

- ### Messages
  Email, SMS, AOL IM, MSN IM etc.

3343 Peachtree Road, N.E., Atlanta, Georgia 30326 | www.psydex.com | info@psydex.com | 404-551-5238

PSYDEX AG - PQL (PSYDEX QUERY LANGUAGE)

PQL is an intuitive syntax for requesting information from AG. PQL expressions are made up of words, phrases, symbols and logical operators that define a high order concept (e.g. person, place, thing, and event) as well as more abstract concepts (e.g. fear, sentiment). Multiple languages (e.g. English, Chinese, and Arabic) can be combined into a single expression. In addition to simple word, phrase expressions, the following operators may be combined to produce high-precision queries. PQL expressions can be changed dynamically and are always processed Ad Hoc in real time across all data in AG. Topic Models in AG are simply saved PQL expressions. PQL expressions are used as criteria in API Calls to retrieve Content, Statistics, Time Series, etc.

( )

Sets

Phrases separated by commas where the comma denotes **disjunction (OR)**

ex. (Apple Computer, شركة ابل كمبيوتر,苹果电脑, iPhone, iPod, ipad, imac, itablet)

+

Conjunction (AND)

Co-Referenced within 1024 tokens or 10 seconds in a Stream

ex. (IBM,آي بي إم,IBM公司, big blue, International Business Machines,آلات التجارية الدولية,国际商业机器) + (acquired,مبستكم,收购, is acquiring, will acquire)

{n}

Unordered Proximity

Tokens are adjacent to each other and within a specified # of tokens apart

ex. (IBM,آي بي إم,IBM公司, big blue, ,آلات التجارية الدولية,国际商业机器,International Business Machines) **{5}** (acquired,مبستكم,收购, is acquiring, will acquire)

[n]

Ordered Proximity

Tokens are adjacent to each other and ordered within a specified # of tokens apart

ex. (IBM,آي بي إم,IBM公司, big blue,آلات التجارية الدولية,国际商业机器, International Business Machines) **[5]** (acquired,مبستكم,收购, is acquiring, will acquire)

$

Semantic Expansion

Expand topics based on symbol and morph expression to include related terms.

ex. ($IBM) {10} ($Acquisition)

-

Negation

Match where include terms are present and excluded are not

ex. Iraq – nuclear

3343 Peachtree Road, N.E., Atlanta, Georgia 30326 | www.psydex.com | info@psydex.com | 404-551-5238

PSYDEX AG - SOLUTIONS (Powered By Psydex)

The solutions below were developed using AG's robust APIs which include REST and Java SDK. Custom TCP/IP socket interfaces are also available as well as client samples for Java, C# and Adobe Flex.

AG for Defense/Intelligence

Security interests around the globe are challenged to respond to a wide variety of real threats ranging from Network Attacks to Information Operations (IO). Time critical threats can emerge in seconds and analysts cannot afford hours or days to identify and respond. AG adapters for Network Traffic and a wide range of News, Email and Social Media can consume and analyze billions of messages a day to identify unusual patterns in milliseconds to alert analysts. Psydex Social Surveillance tools and advanced analytics can filter out the noise and detect real events by analyzing normal levels from multiple sources and correlating across streams.

AG for Trading

Trading is global and cross-class trading in equities, commodities and currency markets is the norm. As High Frequency Trading approaches ZERO latency traders are seeking alpha and insight from alternative sources such as News and Social Media. Psydex has developed Applications and tools for Traders under the name Psyng ("sing") and are available through Web based tools and Instant Messaging such as AOL Instant Messaging. Psyng combines and correlates information on publicly-traded companies with market data and news from many diverse sources, including major News Wires, Social Media, and Television.

AG for Regulatory/Compliance

Regulators today have the overwhelming task of performing surveillance of market activity across multiple exchanges, classes of instruments and accounts. AG fuses market data via feeds from major exchanges to analyze trades, and correlates trades with news and market moving information disseminated via major News wires, Television, Social Media, Blogs and proprietary sources. The solution leverages Complex Event Processing (CEP) to identify suspicious trades when compared to configurable metrics, such as VWAP, where deviations occur ahead of unscheduled news releases. These trades are identified and can be further investigated with market participants and regulators.

AG for Advertising

SEO Advertising firms are looking ahead to the "Holy Grail" where smarter Ad bidding factors into the analysis of ***Real Time*** current events, sentiment, and other factors that affect how ad campaigns are delivered. Often times, companies are slow to adapt their advertising because they can't pattern and trend what people are "Thinking" to enable marketing campaigns to be modified or new campaigns created in real time.

3343 Peachtree Road, N.E., Atlanta, Georgia 30326 | www.psydex.com | info@psydex.com | 404-551-5238