Analytic Technology™ Industry Roundtable

Analytic Technology Industry Roundtable Study:

Classes of Analytics

Dr. William Niehaus, CSRA-NetOwl

Mr. Brian Adams, SAPNS2

Mr. Steve Panzer, Centrifuge Systems

Mr. Lorne Hanson, Centrifuge Systems

February 2017

Analytic Technology™ Industry Roundtable

This page intentionally left blank.

Introduction

Over the years, there have been many different types of analytical products built by commercial software vendors for both public and private sector organizations that address specific analytical requirements and challenges. Some products provide generic capabilities and platforms for analytics, while others address specific analytical functions which are often components of a larger, multi-vendor application. This paper focuses on the classes of individual analytics that are available in the commercial marketplace and that have wide applicability to the analytical requirements of different public section use cases. For each of these "archetype" analytics, we provide an example or two from the Use Case Study[1] and how each archetype may contribute to an overall solution. Care has been taken to not provide specific vendors for each capability as, maintaining a list of all relevant vendors is impractical, and the optimal provider/product for any specific application depends on more specific requirements too detailed to enumerate here.

General Analytical System Requirements

Before introducing each analytical archetype, we first describe some common components of all analytical systems.

Data Harvesting / Data Extraction

Managing the content relevant for any analytical application is a common first step in any use-case and analytical workflow. Content can come from any number of different sources, but can generically be classified as unstructured or text-based content, structured content, and "everything else" which includes among other things, video, audio, and biometric data.

The data may already be resident inside the organization, it may be publicly available via internet, or it may originate within partnering organizations or agencies. The data may be provided in real-time (or nearly so), it may be the product of periodically executed batch jobs, or could be made directly accessible by the parties responsible for the originating system's management. The destination of this data is most commonly a data management system that will both host

[1] Etches, Brown, Stultz, 2016. MITRE Roundtable Working Group, *Analysis Use Case Roundtable Study*.

acquired data and/or provide some manner of integrated access to data that cannot be otherwise imported into the system for reasons such as volume or security.

At the commencement of a project there are well known, relevant, data sources. However over the lifetime of these projects new sources will become available. Many will become consistent contributors to the analytics process, where some may be sources of 'opportunity'. The expectations of data quality, consistency, and timeliness delivered as the product of analytics will drive the characteristics required during the ingest processes

A functional example of data harvesting is contained within the associated case study "Identify Counterfeit Products"[2] identified herein as the Counterfeit Product Project Office scenario. As a precursor to any analytical efforts, relevant data and their sources must be identified. Data such as product names, information regarding partnering retailers and descriptive information is contained within internal systems. To determine those who may be offering counterfeit merchandise, due to the concerns with internet 'retailers', external data must also be acquired. Examples of available external data may be in the form of retail websites, auctions, or social media forums.

The sources of data thus have been identified and now the method of harvest or ingest may be considered. Internal system may be replicated to an independent data management platform or otherwise be made directly accessible. Alternative techniques must be considered when determining the acquisition or access to identified external data. Website content for example can be acquired via web-crawling techniques which use keywords, phrases and links to traverse the Internet and extract content. Social media may be accessible via the same web-crawling scenario, or more advanced real-time capture techniques may be employed. The acquisition of data, internal and external, structured and unstructured, binary and human-understandable, now makes it possible to choose and apply the type of analysis most appropriate for the data available in conjunction with the organization's analytical objectives. The desired analytics can help determine the appropriate data repository or repositories that will be necessary to achieve the desired capabilities and performance.

[2] Etches, Brown, Stultz, 2016. MITRE Roundtable Working Group, *Analysis Use Case Roundtable Study*..

Data Repository

Regardless of the source and type of the data described above, such data needs to be captured and relevant parts stored in a data repository. There are many different types of repositories including: document management systems, relational databases, and NoSQL databases among others. These systems may live on individual servers, clusters of servers, or in public or private clouds. This paper is focused on the prototypical analytics that comprise an overall system, but it's important in the overall planning and design of any particular solution to select the right data repository for both initial and projected analytical requirements.

Archetype Analytics

The following sections describe some of the major classes of analytical capabilities providing brief examples of how each applies to some common use cases that have applicability across many different applications.

Text Analytics

Text analytics exploits unstructured or narrative content with the goal of producing key structured information about that 'document' as output. The outputs of text analytics are usually represented in standards-based formats including XML, JSON, RDF, and others, and can often be combined with other structured data sources and document-level metadata in downstream analytics of various kinds. Indexing the results of text analytics can also support knowledge discovery by allowing end-users as well as other analytical processes to request information about a specific *TYPE* of entity in addition to information about a specific entity of interest.

The following table lists major analytical functions that fall in the text analytics discipline.

Function	Description
Entity Extraction	Identification of mentions of people, places, organizations, and other entity types from unstructured content.
Relationship Extraction	Identification of semantic relationships between entities based on context. For example, "is parent of", "is subsidiary of", "is affiliated with"
Event Extraction	Identification of activities and their participating entities. An "attack" event might identify a "victim", an "attacker", a "weapon", a "place of attack", and a "time of attack".
Sentiment Analysis	Identification of sentiment expressions reported either at the "document" level or associated with specific entities mentioned in the text.

Topic Tagging	Assignment of topic of a document based on presence of key words/phrases or combinations thereof.
Document Categorization	Assignment of one or more predefined categories to a document based on a variety of rule-based or statistical analysis of the content. Often involves training phase to teach the system.
Document Clustering	Automated grouping of documents based on similarities identified while processing each document. The number and scope of "categories" are typically not known a priori.
Geotagging Text	Assignment of latitude-longitude values (or other coordinate expressions) to geo-locatable entities identified in free text.
Document Summarization	Identification of key passages of text document, often influenced by specific keywords/phrases used to identify the document via search.
Author Identification	Identification of who wrote a text based on variety of linguistic characteristics of the text. Also can identify likelihood of same person authoring different texts, even if specific author is unknown.

Table: Common Text Analytics Functions

Continuing with the Counterfeit Product Project Office scenario, the unstructured data sources like the collaboration-related discussions harvested from social media and/or forums/discussion groups can be run through various text analytics services to automatically identify key information such as the names of people, places, addresses, or other relevant entities including geospatial information gleaned from the mention of places in the unstructured output. Using both the metadata of the content's origin (*e.g.*, site, screen-name, posting date, etc.) and the semantics contained within the content itself, relationships between extracted entities can be identified. The resulting structured data produced through text analytics can be combined with other downstream analytics to provide the analytics/investigators, a more complete situational awareness.

Structured Descriptive Analytics

Descriptive analytics is usually the first type of analysis an organization uses to understand their own performance and as a result they are generally well understood. Key analytical functions falling in this category include[3]:

1. Standard reports and dashboards: What took place? How does it relate to the organization's blueprint?
2. Ad hoc reporting: How many? Where?

[3] Akerkar, Rajendra, 2013. Big Data Computing, p 378. CRC Press, New York.

3. Analysis/query: What is the challenge? Why is this happening?

Descriptive analytics can be viewed as the means to cull through large accumulated data sets ("Big Data") to assemble associated information. The problem can often be viewed as the accumulation and analysis of structured data from independent systems. Although this data is often tabular in its construct, the data in its original form can also be unstructured and refactored through processes such as text analysis such that even data originating as unstructured can be presented to the user community in a tabular form. Descriptive analysis endeavors to utilize the relationships between independent systems to provide decision ready information. Job descriptions may be compared with resumes to identify qualified candidates. Human Resources must be associated with learning/knowledge management to appropriately assign personnel, supply chain management, asset management, maintenance history and perhaps even geolocation information must be correlated to determine what parts are necessary to keep assets up and running.

Descriptive analysis provides information relevant to a given situation and will help answer questions in the moment: "Do we have a person who can perform the job"? "Do we have the parts necessary to carry-out the required repairs"?

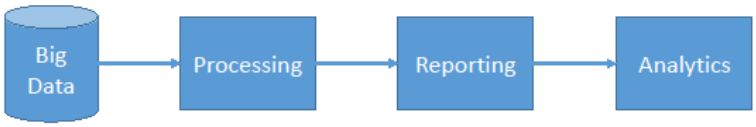

Image 1: Descriptive Analytics Value Chain [4]

Extending the Counterfeiting Product Project Office scenario we have the goal of not only identifying potential violators but discounting those who are legitimate vendors of the products. To identify legitimate vendors we must answer the question: "Does this site have the authority to offer the products they display?". Through the acquisition of web content and the subsequent text analysis we will have established the qualities of a given retailer, among them site address, and

[4] Akerkar, p 377

products on offer. We can compare this information with internal information regarding valid retailers to ascertain whether further investigation is warranted when the site or product is not contained in an existing agreement.

Predictive Analytics:

Where descriptive analysis allows an organization to provide situational intelligence, predictive or anticipatory analysis will use the known behaviors of the past to determine the likelihood of an event, or series of events, occurring in the future. Predictive analysis is the application of data science and statistical algorithms to determine the probability of these future outcomes. Predictive analysis is intertwined with the promise of "Big Data". How can an agency's or organization's retention of large volumes of data be effectively analyzed to deliver a forecast with some acceptable degree of confidence.

Asset maintenance is an example of where the application of predictive analysis has been applied successfully. Applying predictive algorithms to historical information, asset maintenance becomes predictive asset maintenance. Information about the reliability of assets, including capital resources (*e.g.*, vehicles, machinery, etc.), and infrastructure (*e.g.*, antennas, pipelines, etc.), can aid in the delivery of new policies and procedures to maximize uptime and utilization. Predictive Analytics may forecast that in desert operating environment a specific vehicle is expected to operate without problems for 5,000 miles. While there is no assurance that at 5,000 miles the vehicle will cease operation, it does provide the organization valuable information to preemptively apply maintenance procedures to limit the likelihood of operational interruption. With the information provided in this basic example, one can see that the power of these predictive conclusions can not only improve maintenance policy and procedures but be can be used to alter upstream operations including improvements in the design and manufacture of a given part subject to failure.

Predictive analytics implies that we have the ability to build and compare predictive models (regression analysis, decision tree, neural network…to name a few) that use historical to determine the probability of future events. With the predictive maintenance example, the historical data captures real-world reliability information to help predict future failures. In some cases however, the actual outcomes you are trying to predict have never occurred or are rare enough to not provide enough information to, by themselves, meaningfully help predict future related events.

In the absence of known examples of misconduct, which is likely the case in the Counterfeit Product Project Office scenario or in another example, *i.e*., the Money Laundering scenario[5], we can treat the data as "unlabeled" which means we don't have known instances of that we are trying to discover and are not able to build predictive models but can use "unsupervised learning" techniques that do a good job of automatically discovering patterns and trends through clustering. The clusters that are created will often identify the groups of behavior (popular in number) that we can examine and identify as "normal behavior" and perhaps groups of behavior (low in number) that are suspicious. In the Counterfeit Product Project Office scenario, the manufacture level data we have collected through our data harvesting would likely produce clusters of those that are behaving similar (multiple products being advertised, longer dates of selling products and being in business as compared to manufacturers with only have one product with a short period of the manufacturer being in business. In the Money Laundering scenario, clusters of related financial transactions that differ from the target's other transactions and that are different from clusters identified in the transactions of other similar businesses that are less suspected of illegal activities, can draw attention to specific transactions and behaviors that can investigated more completely.

Geospatial Analytics

Geospatial analytics leverage a variety of algorithms and techniques to discover or provide insight into spatial awareness. Examples of geospatial functions include map generation, alignment of spatially-oriented imagery to maps, geographic-feature/surface analysis, and visualizations of geographic data obtained through other analytics types including text, structured data, and imagery analytics. Overlaying geospatial information across a temporal dimension also can provide key analytical insights to many different challenges.

With our Counterfeit Product Project Office scenario and the output from our geospatial clustering, we might find that there are significant geolocation elements in the data that help define the separation in the clusters. For example, those manufacturers who are licensed as a business in the U.S. might be cross-referenced with production facilities overseas while an anomaly might be found where one manufacturer that is licensed as a business in the U.S. and producing the products from the same address. This is not typically what we see and might warrant further scrutiny to see if this type of product is legitimate.

[5] Etches, Brown, Stultz, 2016

Other applications of geospatial analytics capabilities can be found in the Money Laundering scenario and in another example, *i.e.,* the Terror Finance scenario[6]. In this latter case, by tracking down the middle men and the bank transfers, these can be plotted and linked on a geospatial backdrop to uncover a global terror finance network and how it operates.

In yet another scenario, *i.e*., the Financial Crimes scenario[7], geospatial clustering can help identify a bank network that may be enabling money laundering schemes. In network security, organizations offer IP location services that can identify where an IP address is geographically located.

Geospatial analytics also provides an example of a specific analytical discipline that has invested and is now benefiting from the development of key standards that facilitate interoperable GIS-related functions. The Open Geospatial Consortium (OGC) was started to develop and evolve these standards which are now supported by nearly every GIS platform. It's important to note that tools may use whatever data formats desired within the boundaries of its own system, but they are now expected to support the production and ingestion of OGC-standards-based content. Other analytical disciplines may be able to benefit from a similar community-wide investment.

Link/Network Analysis

Most people understand the use of relationship graphs or link analysis when applied to social networks, telecommunications networks, or financial networks. But this analytical method has application in a wide range of use cases where the relationship between entities is crucial to understanding the data, whether the data is structured or unstructured. In addition to traditional intelligence applications, link analysis can be applied to cyber security, healthcare infection/disease tracking, fraud detection, and insider threat monitoring.

The visualization of relationship graphs provides analysts with additional insight by allowing them to ask questions of the data that are not easily represented in traditional BI graphs or by SQL queries. Non-obvious relationships can often be

[6] Etches, Brown, Stultz, 2016

[7] Etches, Brown, Stultz, 2016

detected by analyzing large datasets down to entities of interest, and then expanding out from these nodes to see their 1, 2, 3 or "n" hop relationships.

In the Counterfeit Product Project Office scenario, using link/network analysis, we can visualize the connections that might be too complex to understand within a table of data but presented as a visual network, we can quickly understand the relationships and patterns that would help us identify suspicious behavior. For example, manufacturers typically promote their products in social media and have a long history of content as product lines change (by season) and styles go in and out of favor. If a manufacturer is producing a counterfeit product and they have an absence of social media content around that type of product, it might hint that this manufacturer is new to the business of selling this or similar items. If this category pops up as new to a product line within social media and internet advertising content it might indicate a new counterfeiting operation has been established to quickly off-load select garments. Link/Network analytics could provide a timeline with known manufacturers selling products and those networks with history show an established track record, while a new player in the market pops up as a network that shows up as an outlier and stand out visually from the norm. It might also reveal an extended network of other products that highlight suppliers or other entities of interest that are part of a previously unknown network of collaborators.

Additionally the Counterfeit Product Project Office scenario illustrates how different analytical disciplines, in this case text analytics and graph analytics, can be integrated. As a website's content is discovered through text analysis, entities, objects, and data points of interest emerge and can be classified. The classification of entities into categories (*e.g.*, person, place, item, etc.), coupled with their attributive properties and inter-relationships, can be the basis on which a robust graph can be built. Entities become the nodes of the graph, relationships its edges, and extracted attributes can be applied to either where appropriate.

Streaming Analytics

Streaming analytics straddles the divide between data acquisition and response. Data streams, such social media, financial tickers, or signals intelligence by themselves are data sources. 'Streaming data' becomes 'streaming analytics' when the discrete elements within an incoming data stream are utilized to determine whether variations fall within boundaries, which may spawn an event or action. Streaming analytics, otherwise known as complex-event-processing (CEP), can be viewed as the means to provide real-time situational awareness and, as necessary initiate action, without active human intervention. Streaming is the highest velocity method of ingest in the analytics spectrum. Streaming

provides continuous analysis of discrete events. These events may be combined with other events, or enriched with data from other systems to be evaluated against an identified tolerance threshold.

Conventional analytics is a user-driven process of query refinement against a pool of data. Each subsequent query's result, allows the user to determine another possible refinement to direct the path of investigation. Each inquiry is altered and initiated by a user.

In contrast, when leveraging streaming analytics the question that is to be 'answered' is predetermined; it is the data that is ever changing. Streaming is applied when the highest level of responsiveness to changing circumstances is critical. Within the financial sector for example, where high velocity trading necessitates real-time situational awareness, organizations use real-time streaming to effectively determine the appropriateness of their equity positions.

Valuable information streams are all around us. Streams are commonly associated with sensors that have been deployed for a specific purpose (*e.g.*, signal buoys, AIS transponders, etc.). However, streaming data and, the value therein, exists ubiquitously in the geolocation available through mobile devices as well as the information that can be acquired within the realm of constantly updating social media feeds.

Streaming Analytics provides another example where the intersection of analytical disciplines is starkly evident. Streaming location data can be integrated with geospatial analytics to determine when a geo-fenced threshold is nearing or has been crossed and initiate an appropriate response. Predictive models can be embedded within a streaming context to determine whether a particular event or series of events falls within the predicted probability of aberrant and actionable behavior.

A good use case for streaming data can be found in network security risk. An organization with established thresholds of activity can use streaming data to alert the ISSO of unusual and risky activity on the network from either inside or outside of the network. This activity typically comes from a network analysis tool that detects port scanning, probing, and multiple hits on the same IP address.

Other Analytical Classes

This section describes some other classes of analytics that have their place in specific analytical architectures and systems. More details on these classes of analytics may be filled in in future versions of this paper.

Imagery Analysis

Imagery Analysis describes the extraction of useful information from any sort of sensor-related data contained in bi-dimensional graphic format. Examples of data include photographs, remotely-sensed multi-spectral imagery from satellites and aircraft, ultrasound, magnetic resonance imaging (MRI), infra-red photography, seismographs, et al. Examples of imagery related analytics include facial recognition, geographical feature identification, optical character recognition (OCR), and handwriting analysis.

Audio Analysis

Audio/Voice analysis describes the extraction of useful information from audible or inaudible sound data. Examples include language identification, speaker recognition, sonar, and echo location.

Biometrics

Biometric analysis describes functions that help identify specific or general information about living things – usually people. Examples include physical characteristics such as fingerprints, DNA, and retina scans, to behavioral characteristics such a person's gait and typing cadence.

Conclusion - Alignment Matrix

The table below summarizes information about which of the above classes of arch-type analytics typically apply for a collection of generic use cases. Whereas above we provided some specific examples and details of where analytics may apply to some specific use cases outlined elsewhere, here, we cross-reference several additional common use cases against the same analytical classes. There are certainly additional use cases that may be added to this table over time, as well as an updated set of archetype analytics. The specific functional requirements for any specific use case will dictate which particular products/offerings are most appropriate for any deployment. Government sponsors who have similar or analogous use cases should enumerate their specific analytical requirements in each area to determine the right collection of interoperable components to meet their needs.

USE CASE\Analytics		Text (Unstructured) Analytics	Structured Descriptive Analytics	Predictive Analytics	Geospatial Analysis	Link/Network Analysis	Imagery Analysis	Voice Analysis	Biometrics	Streaming Analytics
Cyber (defensive and "other")		O	C	O	O	C	O	O	O	C
Insider-Threat / Counter-Intel		M2	C	O	O	C	O	O	M1	M1
Command and Control (or C4ISR)		M2	C	M1	M1	C	O	O	O	C
Law Enforcement		C	C	M1	M1	C	O	O	C	O
Counter-terrorism		C	C	M1	M1	C	M1	O	M2	M2
Counter-proliferation		C	C	M1	M1	C	M1	O	O	M2
Counter-narcotics		C	C	M1	M1	C	O	O	O	M2
Financial crimes and money laundering		O	C	M1	O	C	O	O	O	M1
Disaster Relief		M1	C	O	C	O	O	O	O	M2
Search and Rescue		O	C	O	C	O	M1	O	O	M2
Disease Control		M1	C	M1	C	C	O	O	O	M2

Table 1: Alignment of Use Cases to Archetype Analytical Disciplines

Key	Importance	Description
C	Critical Capability	Some types of these analytics are required for these use cases.
M1	Moderate Importance 1	Not critical for complete success but often useful if capability is economically feasible.
M2	Moderate Importance 2	Can provide deeper situational understanding where needed and economically feasible.
O	Optional	Data for these analytics are often not available or relevant to analytical requirements.