Altair

Connecting the Dots:

How data prep addresses three major challenges for the Intelligence Community and improves threat detection

Overview

There are three major challenges associated with the collection and analysis of data across the government: data overload, dark data and disparate data. All of these challenges can be addressed with the addition of data preparation tools to existing data analytics platforms across the Intelligence Community. With data preparation solutions, the Intelligence Community can operate more cost-effectively, analyze more data faster, expedite threat detection, and break down data silos through intelligence sharing across departments and agencies.

1

Data Overload

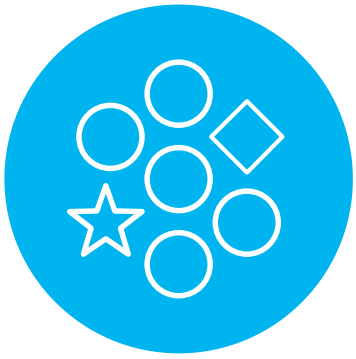
As technology and computer-generated information continue to grow, the amount of data collected by organizations significantly increases

2

Dark Data

Information assets that organizations collect, process and store in the course of their regular business activity, but generally fail to use for other purposes

3



Disparate Data

Each intelligence source collects various data types in different ways, making it difficult to share and combine data from multiple locations

1. Data Overload

First, as technology and computer-generated information continue to grow, the amount of data collected by organizations also significantly increases. For example, each day the United States government collects petabytes of information, but cannot effectively analyze the vast majority of it. GovTech explains: "[g]overnment databases are filled with everything from traffic data to pet-ownership statistics, and many agencies lack the necessary staff and infrastructure to maintain and analyze this information.

Public-sector data analysts report that they spend 47% of their time collecting and organizing data, but less than a third of their time actually gleaning actionable insights from it" (Latham, 2016). More information is good, but when it cannot be analyzed or connected due to volume overload it can be rendered useless. Failing to provide Intelligence Analysts with the necessary tools to leverage all available information, positions government agencies, departments and organizations for intelligence failures.

47% of data analysts' time is spent collecting and organizing data, but less than a third of their time is allocated to actually gleaning actionable insights from it.

2. Dark Data

A second challenge associated with data in the Intelligence Community is the existence of dark data. Dark Data is defined by Gartner Research "as the information assets organizations collect, process and store in the course of their regular business activity, but generally fail to use for other purposes" (Gartner, 2017). Put more simply, dark data is data that is too difficult to bother including in daily threat analysis.

To put this in perspective, Forrester Wave explains that most organizations are only including 12% of all their available data in their analytics. This means that 88% of usable data is being left untouched (Forrester Wave, 2014). For Intelligence Analysts, failing to use all available data from investigative materials (financial records, travel records, social media pages and/or tips from the community) can mean missing key intelligence about a credible threat to the community.

88% of usable data is being left untouched. Most organizations are only including 12% of all their available data in their analytics.

3. Disparate Data

A third major challenge of mass data collection is the existence of data in disparate formats. According to the Director of National Intelligence, "[t]here are six basic intelligence sources, or collection disciplines: Signals Intelligence (SIGINT), Imagery Intelligence (IMINT), Measurement and Signature Intelligence (MASINT), Human-Source Intelligence (HUMINT), Open-Source Intelligence (OSINT) [and] Geospacial Intelligence (GEOINT)" (ODNI, 2017). Each of these disciplines collect various data types in different ways, making it difficult to share and combine data from multiple locations.

Data comes from a wide variety of places, including reports from sensor systems, financial reports, web data, social media data, images, videos and ad hoc tips from the public about suspicious persons or activity.

For example, sensor data generally comes through as a text or log file. Financial reports can exist as a CSV, Excel, TXT or PDF file (Datawatch, 2017). Web data is generally saved in an HTML format. Watch-lists exist in various files and/or databases. Combining disparate, but related data sets can require a lot of manual effort, especially if the data must be re-keyed. This requires a large time investment that many analysts cannot afford to make. As a result, 78% of organizations are occasionally, rarely or never including 3rd party data to enrich their analysis (Dresner, 2017). For security agencies, ignoring difficult data leaves holes in analysis which lays the groundwork for intelligence failures.

78% of organizations are occasionally, rarely or never including third-party data to enrich their analysis.

How Data Preparation Streamlines Intelligence Analysis and Threat Detection:

Data preparation tools streamline data access, manipulation, blending and formatting of data from a wide variety of sources. Simplifying data collection and preparation enables analysts to dig into data and arrive at actionable insights faster. Reducing the grunt work and improving threat detection opens the door for law enforcement to take proactive actions to disrupt malicious plots.

Data preparation ensures that all data is accessible by converting static PDF and Text report data into structured, workable spreadsheets, and enabling users to pull in data from other multistructured and unstructured sources, such as Excel, CSV, HTML, JSON, log files, XML files, web pages and a number of major databases. Once this data is accessible, additional cleansing, combining, blending and manipulating provides Intelligence Analysts with data that is in the optimal format for analysis. Additionally, data preparation tools can be set up to automatically format, organize and appropriately name datasets prior to upload into data warehouses, which helps maintain the orderliness of the platform.

The government already uses analytics platforms as a means of mining and analyzing big data from a wide variety of sources. These systems are exceptionally powerful data warehouses, as well as analysis and visualization tools, that enable Intelligence Analysts to streamline analytics, threat detection and decision-making. The platforms acts as a centralized storage location for all important information. Even though these systems are incredibly powerful on their own, data preparation tools for front-end data cleansing can make government analytics platforms even more valuable for law enforcement and Intelligence Analysts in homeland security. Essentially, these solutions sit on the front end of analytics platforms and act as a filter into these data warehouses.

When data warehouses are effectively organized, end-user analysis will be faster, easier and more valid. In more sophisticated data preparation platforms, models and automation systems prove to be invaluable for feeding the cleanest data possible into analytics platforms. Ultimately, the cleaner the original data is, the better analysis will be.

Why Does it Matter?

The Office of Intelligence and Analysis Strategic Plan: Fiscal Year 2011-Fiscal Year 2018 defines its mission statement as: "[e]quip the Homeland Security Enterprise with the intelligence and information it needs to keep the Homeland safe, secure, and resilient" (ODNI, 2011). This plan also explains that one of the Office of National Intelligence and Analysis' primary concerns is the detection and prevention of terrorist attacks on American soil. In today's world, detection of terrorist threats is nearly impossible without the tools for "connecting the dots" on suspicious persons. Much of this is attributed to a change in terrorist modus operandi.

The Southern Poverty Law Center (SPLC) explains that lone wolf terrorism, or terrorist attacks carried out by one individual, is becoming increasingly popular for all types of terror organizations. SPLC notes: "[i]n an age of instant communications and ever more tightly knit societies, the lone wolf style of attack is vastly more likely to be successful than the kind that was once literally planned in rooms full of men" (Lenz, 2015). A study by SPLC confirms that 46 out of 63 terror attacks since 9/11 (74%) were carried out as lone wolf attacks (Lenz, 2015).

While terrorism will never be 100% preventable, arming analysts with the ability to analyze more data, faster is the best way to improve threat detection across the Intelligence Community. In order to attain the stated goal of the Office of Intelligence and Analysis, analysts must take advantage of tools that make their jobs easier and their efforts more effective.

In addition to improving threat detection, data preparation introduces several opportunities for saving agencies money. Without tools to help automate the collection, blending, formatting and dissemination of these disparate sources, analysts will spend the vast majority of their time just getting data ready to be looked at. Time wasted on redundant data cleansing and preparation ultimately translates into financial losses for agencies.

A 2016 Forbes article describes a study which found that data analysts spend roughly 80% of their time preparing data and only 20% of their time actually analyzing it (Press, 2016). To put this into perspective, Blue Hill Research estimates that Analysts who forego data preparation tools waste roughly $22,000 per analyst per year (Blue Hill, 2016). Not only is it a financial drain, but slow or incorrect analysis in the Intelligence Community can be a serious liability to public safety.

Conclusion

Providing Intelligence Analysts with data preparation tools enables them to address challenges associated with information/data overload, the use/neglect of dark data and the existence of data in disparate formats. By removing these roadblocks in the analysis process, Intelligence Analysts will be able to spend less time fighting data and more time focusing on analysis, intelligence sharing, threat detection and plot disruption.

Works Cited

- Biddle, S. (2017). How Peter Thiel's Palantir Helped the NSA Spy on the Whole World. The Intercept. Retrieved from https://theintercept.com/2017/02/22/how-peter-thiels-palantir-helped-the-nsa-spy-on-the-whole-world/
- Boyd, D. (2017). Week 1 Lecture Notes.
- Datawatch (2017). Information Optimization Keeps Edwards Air Force Base, 412th MXG, Mission-Ready. Datawatch. Retrieved from http://www.datawatch.com/wp-content/uploads/2016/12/casestudies-edwardsAFB.pdf
- Datawatch (2017). MasterCard Improves Customer Service with Self-Service Data Prep. Datawatch. Retrieved from http://www.datawatch.com/wp-content/uploads/2016/04/Datawatch_Casestudy_MasterCard.pdf
- Datawatch (2017). Monarch. Datawatch. Retrieved from http://www.datawatch.com/our-platform/monarch/
- Forrester (2014). Big Data Hadoop Solutions. The Forrester Wave.
- Gartner (2017). Dark Data. Gartner. Retrieved from http://www.gartner.com/it-glossary/dark-data
- Latham, E (2016). Is the Government Hoarding Too Much Data? GovTech. Retrieved from http://www.govtech.com/opinion/Is-the-Government-Hoarding-Too-Much-Data.html
- Lenz, R. (2015). Age of the wolf: A study of the rise of lone wolf and leaderless resistance terrorism. Southern Poverty Law Center. Retrieved from https://www.splcenter.org/sites/default/files/d6_legacy_files/downloads/publication/lone_wolf_special_report_0.pdf
- Lev-Ram, M. (2016). Palantir Connects the Dots with Big Data. Fortune. Retrieved from http://fortune.com/palantir-big-data-analysis/
- Miles, A.D. (2016). Intelligence Community Spending: Trends and Issues. Congressional Research Service. Retrieved from https://fas.org/sgp/crs/intel/R44381.pdf
- Office of the Director of National Intelligence (2017). Office of the Director of National Intelligence FAQ. DNI. Retrieved from https://www.dni.gov/index.php/about/faq?start=2
- Palantir (2017). Palantir Gotham. Palantir. Retrieved from https://www.palantir.com/palantir-gotham/applications/
- Press, G. (2017). Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes. Retrieved from https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#6f637daa6f63
- Saylor, R. (2015). How much does Palantir cost? Quora. Retrieved from https://www.quora.com/How-much-does-Palantir-cost
- Southern Poverty Law Center (2016). Hate and Extremism. Retrieved from https://www.splcenter.org/issues/hate-and-extremism