Data in the Aggregate: Discovering Honest Signals and Predictable Patterns within Ultra Large Data Sets

K. Lossau, J. Larson
Sotera Defense Solutions, 1515 S. Capital of Texas Hwy; Austin, TX 78746

ABSTRACT

Traditionally information fusion has focused on the tactical value of finding and tracing a single needle in a haystack. While this approach provides value, it focuses only on a single person instead of identifying the entire culture, community, and scope of a target organization. Data analysis in the aggregate can provide immense strategic value, especially in identifying honest signals[1] and habits (often unintentional). Aggregation of data through data warehousing has been used on large data sets to enhance query response times by summarizing or partially summarizing the data over various dimensions (e.g. pivot tables) or grouping data based on relationships (e.g. clustering). We continue to explore how to use data aggregations as additional data elements to be further processed, analyzed, and queried. We will discuss several mechanisms for analyzing different types of large data sets including dimensional databases and graph data through the application of cluster computing (both in memory and file based representations). This strategy will employ several information fusion techniques that operate on these aggregations to detect anomalies, discover correlations and present historical patterns within the datasets. Approximation techniques, which can be used to reduce the computational order of complexity, are also discussed.

1. ISSUES ANALYSTS FACE TODAY

Data continues to be generated and digitally archived at increasing rates, resulting in vast databases available for search and analysis. Access to these databases has generated new insights through data-driven methods in the commerce, science, and computing sectors. Senior DoD leaders have said the Defense sector is "swimming in sensors and drowning in data.[2]" The so-called "big data" problem has now become a challenge for military operations, both at strategic and tactical levels. The data being brought to bear on operations are growing rapidly in volume and complexity, and are most often imperfect, incomplete, heterogeneous, and consumed by diverse end-users from analysts to field soldiers. Defense applications now have environments where data can sometimes be seen only once, for milliseconds, or can only be stored for a short time before being deleted. The trends are accelerated by the proliferation of various digital devices and the Internet, which are being used by adversaries in all stages of threat production, from planning to logistics to resource movement to operations. Therefore, it is critical to develop fast, scalable, and efficient methods for processing and visualizing data that not only support ingestion and transformation but also enable fast search and analysis. Aggregations and data projections are important not only for analytic execution, but also for visualization and statistical presentation of patterns with in the data.

[1] "Honest Signals - How They Shape Our World" Alex Pentland, MIT Press, ISBN: 978-0-262-16256-2, 2008

[2] We're going to find ourselves in the not too distant future swimming in sensors and drowning in data," said Lt. Gen. David A. Deptula, Air Force deputy chief of staff for intelligence, surveillance and reconnaissance. going to find ourselves in the not too distant future swimming in sensors and drowning in data,": LTG. David A. Deptula, Air Force deputy chief of staff for intelligence, surveillance and reconnaissance.

Using these base technologies alongside existing feature extraction / enrichment analytics can allow for a flexible platform for the development and implementation of new capabilities that work at scale. In this manner, the raw data is used to build analytic projections recursively, such that each new projection is built off of an existing set of projections. Effectively, this creates a serendipitous ad-hoc development environment, in which many analytics can be tested forensically across historical data, while also enabling a pipeline for pre-defined analytics that are optimized for real-time operations and feedback.

The benefit of these techniques is to provide a rich basis for analysis of the large-scale multiINT environment. The use of a large data framework will have significant advantages:

- Scalability – ability to work at scale in a cloud-type environment by use of map/reduce, bulk synchronous parallel, aggregator trees, and other techniques
- Reliability – ability to partition the data provide quality of service through distributed hardware fault tolerance
- Adaptability – ability to provide new and data aggregated data products
- Modularity – ability to bring new analytic results across a non-static dataspace, and the ability to persist those findings back into an architecture for provisioning and dissemination (and follow-on analysis)

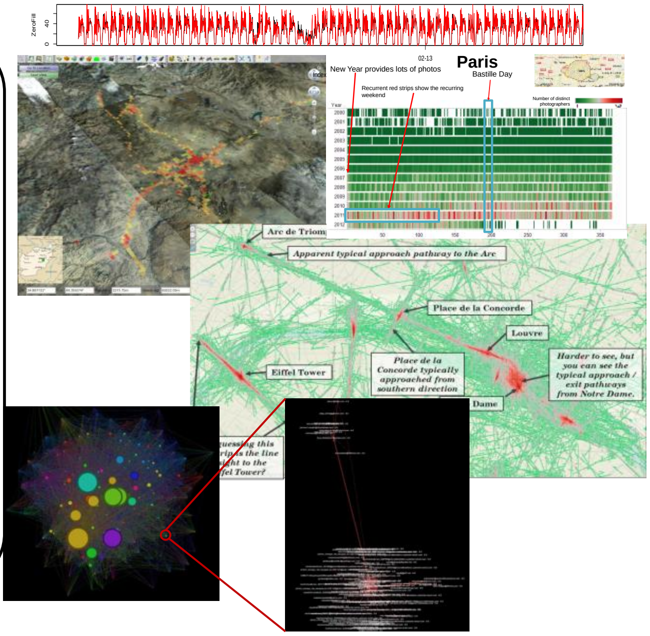
These techniques provide the platform on which analytics can be built to "see what might otherwise not be seen". The ability to pivot and project data from this stance of of deep multi-dimensionality can allow discovery of trends and traits, including those the opponent may not even realize they have.

2. EMERGING DATA PRODUCTS

Large scale data pose significant challenges in locality, transmission, and the partitioning of the data. Sotera and other organizations have increasing more experience in processing these large data streams by using open source cloud technologies, such as HDFS, MapReduce, Hive, Impala, Shark, Giraph, Storm, and Spark which distribute the computational and storage load into a parallelized and scalable environment. For real time or iterative analytics, usage of in-memory distributed systems such as Spark / Shark are used. For pure real time analytics, Storm and Spark Streaming provide excellent capability. For graph processing, Giraph, Bagel, and Hama each provide for a distributed Bulk Synchronous Parallel implementation. . The result of using open source and Government developed software these systems provide users with sub second response to queries to all of the data. Using the projections of the data and the aggregated data products, the analyst can see all of the information and ad-hoc drill down into the relevant information quickly. In the past, analysts were restricted in their query options and often forced to choose selective criteria. This selective criteria biased the returned dataset as it tossed out information that could potentially have value that the analyst had not considered.

The discovery of honest signals in the aggregated data can be discovered by statistically looking at these patterns in the data. Normalizing the data over a common set of features (e.g. time and location being obvious ones) we can uncover patterns with single or multiple data sources. The figures below demonstrate some of these examples. A time series analysis of flight data, traffic patterns in a region at a specific time, social media displayed on a map, and large graphs aggregated through hierarchical clustering with a drill down capability.

- Large Multi-Int analysis
- Data mining, discovery, advanced search
- Discovery, trend analysis and correlation
- Modeling, entity management and Pattern of Life
- Multi-INT correlation across disparate sources



Drawing on experience and techniques that have been leveraged and refined in the multiple DoD programs, analytics can be built for combinatorial calculations en masse by pair-wise combinations across many different data sources and dimensions at scale. Further, approximations techniques can aid in discovering highly correlated items and greatly expedite algorithms that would otherwise have computational complexity that is intractable. This capabilities together formulate an automated correlation capability that can uncover non-obvious relationships between datasets that would otherwise be left undiscovered. The initial results of these analytics, which range into the trillions of comparisons, have provided groundbreaking results that prove the worthy of applying massive computational power to discover otherwise hidden patterns. Sotera has also developed several classes of statistical anomaly detection algorithms, aggregate movement characterization analysis, and rasterization techniques that are optimized to operate at ultra-large scale and can be applied to new problem sets. These analytics are given a final polish through visualization using a set of multi-dimensional and graph tools tailored to analyst's needs.

3. CONCLUSION

Analysts will increasingly be looking at aggregated data products consisting of multiple sources of data fused together to provide an understanding of normal patterns of behavior. In addition to looking at patterns-of-life over individual people, specific entities, events or locations, patterns-of-life products can be created on behaviors and other dimensions of communities, regions, large corporations, and ethnic, religious, cultural organizations. New information can be compared to known trends to discover anomalies and changing patterns of behavior. Where previously the job was to filter through large sources of data to find specific pieces of information that fused together tell a picture, now it is the large data itself that is the product and when fused on different levels reveals patterns and trends within a given slice of the data. The challenge is in finding the right people to excavate the relevant dimensions within the data to create meaningful and relevant aggregated data products.