Data topography for pervasive, proliferated space situational awareness

Phillip M. Cunio
ExoAnalytic Solutions, Inc.
Ben Corbin
IDA/STPI
Brien R. Flewelling
ExoAnalytic Solutions, Inc.

1. ABSTRACT

As the importance of Space Situational Awareness (SSA) for both national security and commercial enterprise continues to increase, more attention will be paid to collecting SSA data by national, corporate, academic, and amateur actors, resulting in ever-growing volumes of data available. In addition, the ongoing development of proliferated, automated collection networks will create growth in the collection rate per network, generating overall an exponentially-expanding amount of available data.

This data volume will need to be transferred, stored, organized, and catalogued in order to be of maximal utility to stakeholders in the SSA domain. However, managing a large data volume can be challenging, particularly when the data sources are multiple providers with varying levels of capability and tasking criteria. This paper addresses a few of the challenges of dealing with large volumes of SSA data, by examining the inherent features of 'lakes' (general open storage constructs) of data, including the effects of inadvertently incorporating low-quality or superfluously-duplicative data, and comparing them to the features of 'waterworks' (pre-routed and extensively-labeled warehouse constructs) of data.

The differences between lakes and waterworks are examined in terms of estimated relative costs to instantiate, maintain, and operate; probability of being amenable to various classes of stakeholders in the SSA domain; and general deployment challenges. This paper also investigates the probable rate of growth in available SSA data, and assesses the scaling of data lakes and waterworks that may result if said available data is aggregated rather than being left partitioned and usable only by its original collectors. From this, a general assessment of the value of aggregated data is made, and recommendations for future data architecture guidelines are given such that a wave of SSA data will not accidentally swamp the capacity of future users to digest it and extract valuable meaning.

2. BACKGROUND FOR SSA DATA

The criticality of Space Situational Awareness (SSA) data for the economic and national interests of the United States and the world is high and likely only to intensify. A substantial fraction of people in the world touch data that has transited through space, been generated in space, or been derived from space on a daily basis. The literal global reach and high ground provided by space (particularly by the geosynchronous belt) also represent a regime of key military advantage for all US service branches.

The availability of deep and rich datasets, however, provides particularly key insights into the behavior of objects on orbit. As described in [1] and with additional context from [2], full-sky persistent coverage provides early warning of pending events and staring (rather than periodically sampling) an ongoing event can provide data of key interest for forensic analysis. The forensic analyses presented for AMC-9 and Telkom-1 in [1] and [2] were of sufficient depth to allow for the generation of initial hypotheses as to precise timelines and events, and the eventual close assessment and rejection of some of these, leading to the conclusion that Telkom-1 most likely suffered a short-duration, high-energy event leading to a catastrophic fragmentation – consistent with a tank failure during a maneuver burn. The subject of the other incident, AMC-9, can be shown probably but not definitively to have avoided a collision with a small, fast-moving but invisible object, and showed evidence of a rapid but changing rotational rate in the weeks leading up to an apparent partial breakup, pointing at (most likely) an attempt to recover from an initial failure that resulted in unforeseen complications.

These concrete examples show the value of rich datasets, including their ability to support pattern-of-life analyses. However, it is something of a misconception to believe in data lakes as the source of rich data. While it is doubtless the case that a data lake probably contains valuable data, and possibly in rich density and depth, a data lake is also vulnerable to certain other effects.

For this work, we define a data lake as a set of data, usually stored in one architectural location (such as within a single server bank or on a connected cloud network), relevant to a given domain of research and operations, sourced from any provider willing or compelled to contribute and updated on a frequent basis. This necessarily implies a very large amount of the data collected within that given domain of research or operations will be contained within the lake. Note that [3] defines a data lake as storing data regardless of format and implies across a very broad range of sources.

As noted in general and in specific example by [4], utilizing a data lake requires more than simply collecting data, and larger data sets do not necessarily mean better results.

3. ISSUES RELATED TO DATA LAKES

Placing all available SSA data into a grand lake and making it available with no further comment is somewhat akin to turning a linguistics department loose in the Library of Congress and asking them to do some good research. While it is not a bad idea, it ignores the underlying inherent assumptions, does not set focus carefully, and overlooks the way what is being studied is actually used.

Data management and classification may become a problem if data lakes are used. The use of a data lake, wherein all data sources pertinent to SSA are combined into one extremely large body, has two potential negative effects. The first is the possible negative effect of low-quality or inaccurate data, a small amount of which may contaminate larger sets of better data (the lake analogy is perhaps apt here, as the negative effects of small amounts of pollutants upon water have been attested since antiquity and publicized since at least 1962, wherefore we may perhaps call this the Carson Effect, to which data lakes are vulnerable).

The other is the dilution of valuable data by other data, a factor to which an SSA data lake (by the nature of the concept, including data amounts exceeding the ability of any single entity to handle, process, and utilize efficiently) may well be particularly vulnerable. If a data lake includes substantial amounts of data from multiple sources, it is likely to include multiple duplicate perspectives of less-interesting events, and the simple effect of sorting through all the data thereby encompassed to arrive at a usable depiction of a news-worthy occurrence may well consume sufficient time and resources that the data lake will actually serve to reduce the value of good data by dilution, thereby achieving the precise opposite of its intention. We may term this the Faelivrin Effect, after the literary description of an event wherein a valuable opportunity to prevent a catastrophic loss was missed due to inattention in the face of a dangerously distracting preoccupation, much to the later detriment of all concerned. Equally may we suggest that the pursuit of a data lake concept, while motivated by appropriate concern, is perhaps an *ignis fatuus* given the resources and attention it would consume in the face of a rapidly-changing and potentially quickly-compounding situation at the GEO belt today.

As noted in some depth in [1], while it seems logical that large lakes of data have value commensurately greater than the small ponds or tiny droplets of data currently available to most stakeholders in the SSA regime, it is more accurate to note that a lake has much greater complexity than does a pond or a droplet. More data may be better, but there should be distinctions made between data from all sources globally and from well-understood, well-managed sources.

4. CURRENT APPROACHES TO DATA LAKES

Ref. [3] particularly acknowledges the inherent problems of large data lakes; in its suggestions for applying organizational structure (of the minimalist style suggested by the term ‘data lake’), the work might be more accurately characterized as describing a metadata lake, which is a repository of carefully-designed references to the deeper, less-organized data contained in the actual data elements (personal browsing data) of interest in the paper. It may also be noted that the significant issues identified in this paper address only the data stream that a part-time

activity of a single human may produce, not a full-time distributed effort by multiple well-funded corporations or governments, such as SSA data may constitute.

Ref. [5] elucidates that data lakes do not provide schema for a massive body of data until the point of use, and specifically notes the somewhat-uncharted legal territory opened by the use of such a data lake, including the fact that they know of no existing enterprise that successfully navigates these issues with a data lake used for new applications. As a further touchpoint, it is noted that idiosyncratic 'conversational' metadata often serves a crucial role in managing large datasets; by design, a data lake has no native format for such metadata types nor a facility to engage the various stakeholders in such a data lake as to allow unofficial channels for such metadata to be handled.

Part of the state-of-the-art in SSA data management involves the use of the Resource Description Framework (RDF) Triple construct, described in more detail for the relevant domain in Ref. [6]. An RDF Triple is an adapted tool from the domain of semantic web technology, but as noted, it requires a separate process to map intake data into the storage form, and the triple is sufficiently loosely structured as to require functional dictionaries and vocabulary sets (separate forms of metadata requiring additional cataloguing effort and attention). Although the RDF triple is a simplified structure for metadata handling, its existence and proposed use for SSA data points to the need to address the fact that metadata use becomes more critical when data lakes become more voluminous, especially if the data are simply dropped into a lake of indeterminate size.

5. INFORMATIONAL ANALYSIS OF DATA LAKES

The informational structure of a data lake may be modeled as a large body of data elements, all pertaining to a subject of known interest, from which some subsets of the data describe some recognizable or interesting pattern may be drawn. A data lake (or any body of data in general) will contain some number of sets of data that detail these interesting patterns, and these interesting patterns can be associated, in conjunction with appropriate processing techniques and algorithms and with relevant knowledge extraneous to the data themselves, with knowledge of great utility to users of the data. It is these patterns, once made visible to the right users, that are the truly valuable component of data bodies, and it is worth noting for future analysis that this value must be extracted from datasets. Data in itself contains value, but this value cannot be realized until some other information-containing element (often called knowledge or news) has been extracted from the data.

Accordingly, a data body contains within itself some number of sets of valuable information. We may presume that these can be quantified via the following model, where $N_{intsets}$ describes the total number of interesting sets, f_{not} is the fraction of possibly interesting sets which are found to be actually interesting (here treated as a constant), and the combinatoric term $C(N,k)$ describes all possible relationships of *k* items in a group of *N* total items, here used to model the total number of possibly interesting sets within a data body of number of data elements *N*, given the logic that an interesting relationship may be considered to address only some fraction, here sized by *k*, of all data present:

$$N_{intsets} = f_{not} C(N,k) = f_{not} \frac{N!}{(N-k)!\,k!}$$

We may further take as an event worth analyzing the collection of some smaller number of new data elements, whence it is possible that a new interesting relationship may emerge, and describe this additional data set with:

$$N_{addlsets} = f_{not} \frac{N_{new}!}{(N_{new} - k_{min})!\,k_{min}!},$$

where the subscripts *new* and *min* pertain to the new data elements collected and the minimum size of data set in which any relationship may be found. In any case, $k_{min} \leq k$, although the two may be set equal for the simplest cases, and $N_{new} \leq N$.

The most notable case occurs when the additional data collected are patterned into a set which shows a relationship somewhat but not entirely similar to previous datasets identified as interesting. The most direct analogue for this situation in the SSA domain is the case where a specific interesting dataset shows a given satellite's typical pattern of life, and additional data freshly collected shows a noticeable deviation in that established pattern of life. In most

cases, this would trigger additional interest, and the mere ability to recognize the novelty of a just-exhibited pattern of life would constitute a clear value proposition for the enterprise behind the data body.

However, this is a relatively ideal case, and we may posit three basic types of data bodies: plain data lakes, which collect any and all data fed into them, with no regard for source duplication or data validation; filtered data lakes, which collect only accurate data but do not distinguish between various sources providing said data; and data waterworks, which have established data validation and redundancy elimination features.

The model for identifying the interesting case discussed above we may pose in terms of costs encountered when attempting to identify said case, namely the time required to perform a certain number of comparison searches between known interesting datasets and newly-acquired additional datasets. This is not necessarily the most truly accurate representation of costs that can be used to compare the various types of data bodies, but it is intuitively a feasible basis for simple and initial comparison. As such, we may set two parameters: β is the number of interesting sets found in a given data body, where interesting sets include all those sets which appear to be interesting or describe interesting relationships according to known algorithms or techniques; and α is the number of interesting sets found in a (smaller) collection of newly-acquired data elements. A fairly obvious method to estimate the time required for this analysis would simply multiply the numbers, as $t = \alpha\beta$.

However, this time should not scale directly across each of the types of data bodies, not least because, all else equal, the various data bodies have variously lower or higher criteria for including data within themselves, and therefore will be of different sizes in any world which has options for multiple data sources. Table 1 shows relative scaling.

Table 1. Features data sets and cost factors within plain data lakes, filtered data lakes, and data waterworks

Data body	Plain lake	Filtered lake	Waterworks
Interesting sets	$\beta = k_\beta\beta$	$\beta = k_\beta\beta$	β
Additional sets	$\alpha = k_\alpha\alpha$	α	α
Time cost factor	$t = k_\alpha\alpha k_\beta\beta$	$t = \alpha k_\beta\beta$	$t = \alpha\beta$

The time factor associated with k_α may be understood as the consequences of the Carson Effect, much as the factor for k_β may be considered the marker of the Faelivrin Effect, where both k_α and k_β are necessarily greater than unity. The scaling for both is a function not only of issues not treated in this paper (such as data transfer inefficiency within increasingly larger databases and the need for constant conversion among formats for data ingested under disparate formatting), but of the time cost differences associated with performing the setup work needed for a data waterworks and the two types of data lakes (again, some practical IT issues are not treated here, as they more rightly pertain to network engineering). Through some analysis rather out of the scope of this paper, we may note that:

$$k_i \propto k'kf_{not},$$

Where k_i describes any modifying factor on a time cost, and *k* and f_{not} are as described previously. The factor k' is approximately a multiple of a) the ratio of all data element providers to those who provide validated data and b) the approximate duplication factor of data from various providers. It is reasonable to assume that, for present-day SSA, both of these factors fall between 1.5 and 5, making 3 a reasonable estimate of the value of either, and thus setting k' to a value of about 9.

With these estimates in hand, we may also set a few variables to notional values in order to arrive at instructive values. For instance, we may estimate *k* as 10, and f_{not} as 0.1, with k' set to 9, and with a total number of data elements at 10,000 (=*N*), and β set to 1000, with α at 20; we arrive at:

Table 2. Time cost factors and scale estimates

Data body	Plain lake	Filtered lake	Waterworks
Time cost factor	$t \propto (k'kf_{not})^2\alpha\beta$	$t \propto k'kf_{not}\alpha\beta$	$t = \alpha\beta$
Time cost estimate/1000	1620	180	20

Note that all three time cost estimates are linear in β, which itself is closely and directly related to *N*, wherefore we note that the variation may be expected to grow as the data bodies themselves do over time.

Fig. 1 shows the notional distinctions between the plain data lake, the filtered data lake, and the data waterworks concepts.

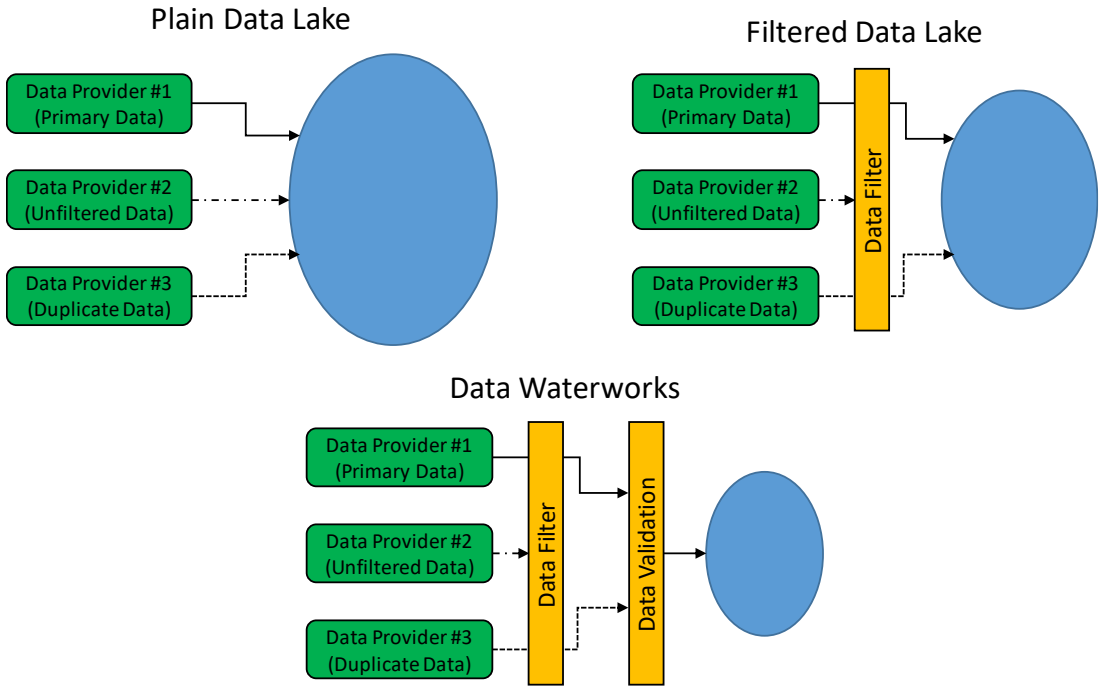

Fig. 1. Notional figure detailing data lakes (plain and filtered) and data waterworks.

Estimating the Size of a Data Lake

In order to understand what a data lake for SSA might look like, it is useful to estimate what the maximum size of such a lake would be based on an ideal state of SSA data collection in the future. There are many different types of data that could be entered into a data lake given the vast differences in data collection technologies. Although one could upload imagery directly, it is more useful to consider the elements being uploaded into a data lake as streamlined observations of space objects. SSA data collection entities can pre-process this data into an observation-centric format rather than uploading raw images to the data lake.

An observation uploaded to a data lake would require more than a simple two-line element. The observation may contain astrometric elements (either in original form or converted into some form of orbital elements), photometric elements, a timestamp, relevant information about the collection circumstances (including sensor type, sensor location, or local atmospheric conditions). Observations collected today may be as small as one kilobit, although in the future it may not be unreasonable to see this rise to ten kilobits per observation, depending on what additional information is necessary for understanding the noise characteristics of a specific observation.

The website *space-track.org* provides orbital data for approximately 19,000 objects today [7]. This number is expected to increase dramatically over the next few years, as several companies have proposed launching large constellations of more than one thousand satellites each [8-10]. These launches will also create more trackable space debris. Perhaps more consequently, the minimum size for an object to be tracked will decrease, so many more small pieces of debris will appear in catalogues. As many as 100,000 objects may fall into a trackable size range in the coming decade.

The desired number of observations made on a particular object per day depends on a number of factors. Ideally, a satellite operator may want up to one quarter of an object's orbit tracked in order more precisely to predict when it may intersect another satellite half an orbit away. SSA companies may make on the order of 1,000 observations per

day on objects in geostationary orbit; objects in low Earth orbit move much faster across the sky, and as many as 10,000 observations per object per day may be desired (and may feasibly be produced if radar data is collected).

Today, several companies and governments generate different types of SSA data, though the number of observations per object, the number of objects each one observes, and the amount of data each company generates varies depending on the number of sensors each has, the type of sensors, and their data processing procedures. In the future, it may be reasonable to assume that ten different providers worldwide are collecting information on each object and reporting these observations to a data lake. Ten SSA providers, each tracking 100,000 objects, gathering 10,000 observations of 10,000 bits, each generates 100 terabits (10^{14} bits) of data per day. If these assumptions are upper limits, and none of this data is thrown out, SSA data collection may generate a maximum of 4.5 petabytes of data per year. For reference, a single picture of the entire surface of the Earth with one-meter ground resolution and 16 bits per pixel is approximately one petabyte, so a data lake for SSA data should not be nearly the size of a data lake used for Earth observation constellations taking multiple images of the entire planet every day. More relaxed assumptions (50,000 objects, 5,000 observations at 5,000 bits each by five SSA providers) would still generate 6 terabits of data per day, or over 285 terabytes of data per year.

However, if raw image data is collected instead of ten-kilobit reduced observations, the data flow rate increases dramatically. Image data can be at least six to eight orders of magnitude larger than reduced observation data, depending on the nature and resolution of the sensors. Under the previous assumptions, if raw observations are seven orders of magnitude larger than streamlined observations, one zettabit (10^{21} bits) of data per day would enter the data lake. This amount of data is approximately equivalent to a one-centimeter resolution image of the entire surface of the Earth with 256 bits per pixel. Even with the more relaxed assumptions stated in the previous paragraph, this amount of data per day is more than three times what was generated on the internet daily in 2017 [11].

If the data ingested into an SSA data lake approaches any of these levels, we may expect the scale problems associated with plain or filtered data lakes to scale consequently quickly.

6. SUMMARY

The term ‘data lake’ has been generally visible as a way to describe one central data repository, into which (it seems ideally) all available SSA data of any type would be placed. While it is appealing to imagine a single oracular interface to all SSA data collected throughout history, which would quickly and effectively answer any question posed it, this will be difficult to achieve in practice. Other works show the general need for a sophisticated intake infrastructure to bolster any large and diverse storage construct.

Ref. [12] highlights some of the complicating factors: security and privacy regulations (and associated structured access permissions), relative complacency due to perceived resource availability (that is, some organizations see pooling data as an end in itself, and regard data access as more valuable than its true valuable adjunct of knowledge access), and the resulting failure to build a knowledge-extraction architecture around a data body while the data body grows.

Part of the fundamental appeal of a data lake is the perception that it “removes information silos [12].” While this is more or less true, it may not be accurate to regard the removal of an information silo as a net positive. An information silo is also necessarily a filter, constructed to sieve out meaningful pieces of information and impose some structure on otherwise-disorganized data. Removal of an information silo simply transfers the costs of information sieving, structuring, and management to the end user; end-user processes notably do not scale in a highly economical fashion with extremely large bodies of resources – they would be, in fact, a perfect inversion of the concept of economy of scale.

A more appropriate way to acknowledge the future interest of the national and global SSA community in data lakes may be to note that plain of even filtered data lakes are perhaps not the most apt metaphor, and to begin discussing data waterworks instead.

7. REFERENCES

1. Cunio, P. M., Bantel, M., Flewelling, B. R., and Therien, B. "Advanced Debris Analysis Techniques Enabled by Rich Persistent Datasets." ERAU STM 2018.
2. Cunio, P. M., Bantel, M., Flewelling, B. R., Therien, W., Jeffries, Jr., M W., Montoya, M., Butler, R., and Hendrix, D. "Photometric and Other Analyses of Energetic Events Related to 2017 GEO RSO Anomalies." AMOS 2017.
3. Alrehamy, H., and Walker, C. "Personal Data Lake With Gravity Pull." 2015 IEEE Fifth International Conference on Big Data and Cloud Computing. Published 21 October 2015.
4. "More Than Sensors" featurette, Analytical Graphics, Inc. homepage. ComSpOC page. Accessed 29 Aug 2018. URL: https://www.agi.com/comspoc.
5. Terrizzano, I., Schwarz, P., Roth, M., and Colino, J. E. "Data Wrangling: The Challenging Journey from the Wild to the Lake." 7th Biennial Conference on Innovative Data Systems Research (CIDR), Jan 2015. URL: https://pdfs.semanticscholar.org/2a24/f587b68a1ef6539b4ed8725dfe76f0ed40e2.pdf.
6. Jah, M. K. "ASTRIAGraph." URL: http://sites.utexas.edu/moriba/astriagraph/. Webpage copyright 2017. Accessed 29 Aug 2018.
7. "Objects in Orbit" Satellite Situation Report via space-track.org, compiled and provided by JFSCC, Vandenberg AFB, CA.
8. https://spacenews.com/us-regulators-approve-spacex-constellation-but-deny-waiver-for-easier-deployment-deadline/
9. https://spacenews.com/boeing-proposes-big-satellite-constellations-in-v-and-c-bands/
10. https://spacenews.com/oneweb-asks-fcc-to-authorize-1200-more-satellites/
11. "Data Never Sleeps 5.0." URL: https://www.domo.com/learn/data-never-sleeps-5.
12. "Drowning in a Data Lake," Peter Arena, ASR Analytics, published on LinkedIn, 25 July 2018. https://www.linkedin.com/pulse/drowning-data-lake-peter-arena/.