

# Some Thoughts on Data and File Analytics

By James Willson-Quayle \*\*

The Department of the Navy (DoN), like other government agencies, has done a fairly poor job of managing its unstructured and structured data. However, thanks to some initiatives currently underway, the situation looks to change. The Data Savvy Initiative is one such example of the DoN recognizing the need to train its workforce to meet the data challenges of tomorrow. The DoN recognizes that there are data skills which are not easily found but are nevertheless needed to advance its mission. Fortunately, the DoN also recognizes that there are hundreds if not thousands of DoN military and civilian personnel who have an interest in, and who may be qualified to receive advanced analytics training.

On the big data/structured side, in addition to a skills gap that must be filled, the DoN needs to develop a coherent organizational strategy. As background, data comes in 2 forms: unstructured and structured. For most organizations, unstructured data makes up 80% of content and structured is 20%, and with the total size of data roughly doubling every 2 years. So far most of the discussion has been on the 20%, big data, side of the ledger. Although this is not intended to be an exhaustive list, I would think that in order to sustain a big data analytics capability the following considerations would have to be met:

1. A plan for finding and/or training people skilled in analytics, which might even include Hadoop or similar-level ability.
2. A reconsideration of who owns the data. This is not as simple as it seems. Just because an office manages a dataset does not mean it owns it – and, by extension, it can prevent others from accessing it.
3. A need to baseline data management practices – which data, who has access, using a modular approach.
4. A determination to what purpose the analytics will be used. Is the data management required for a) disclosure of emergent trends (unknown unknowns), b) identifying disruptive forces, and/or c) proactive vs. reactive activities.
5. An investigation if command capabilities, like those found in NAVAIR, can be leveraged as tool suite for basing capabilities and training requirements. There is no point in re-inventing the wheel.

6. A determination what are the objectives/expectations for data analytics (i.e. do we have a requirement to perform “real time” or fast analytics?).

7. A determination of the roles and responsibilities for those who are involved in a data analytics “program”. Once trained, who do they work for – their own office or do they serve as a resource for other SECNAV components?

I am not just referring to an ability to dive into one dataset, but to take full advantage of the data and tools by extracting and merging it with other data in a separate repository or system (data transformation). Combining multiple datasets is a promising way for the DoN to generate innovation. This would obviously require exposing datasets, so that at a minimum metadata of closed datasets would have to be made visible. This would allow end-users to collect useful metadata and for system owners to respond to data requests that might have utility elsewhere in the DoN. Further, the DoN may need to revise its analytic tool standards as it may be difficult to know which data tools are best directed towards to which purpose (analytics tools should be kept separate from the data to ensure the ability to plug in or exchange). Also in need of revision might be the standards by which data is kept: data in whatever form should be easy to read if at all practicable. If analysts from another command or office are to exploit the data, then data in Plain English should be the norm.

The benefits of a coherent strategy need to be made visibly apparent to all. For example, for those who are end users, data needs to be accessible via mobile devices. And at the senior level, we should be aware that the explosion of data does not automatically transfer into better decision-making. But while we wait for promise of using big data for Command & Control (C2) to be realized, we can use visualization and analytics more in briefings to senior leaders or even in senior leader iPads in lieu of traditional trip binders. Such dashboards would be easier to understand, faster to update, and give senior leaders a glimpse into the art of the possible.

Finally, to maintain a strong posture for the future and ready ourselves for the explosion of Internet of Things (IOT) data, we need to be able to process data at the speed that it is collected. On the unstructured side – that is, the 80% of our data that is often found in Word and PowerPoint formats – file analytics tools are available that can reduce this unstructured pile by 40-60%. That would reduce content that clutters our repositories that is redundant, obsolete, and trivial (ROT). File analytics would then apply metadata to remaining content for easy identification, search, and access. The US Army has used file analytics to comb through 60TB of its Iraq and Afghanistan war records, and USCENTCOM had done the same for 200 TB of content it has brought back from the front. The Joint Staff has used this tool to eliminate 80% of the content from its shared drive, and the Marine Corps is currently engaged in a pilot program to provide a semantic search of 2.5 TB of unstructured data from Iraq and Afghanistan.

The semantic search for unstructured data is more than a key word search. In a semantic search, phrases, terms, and concepts are highlighted that might otherwise go unnoticed. A few years ago, a semantic search of After Action Reports buried in the U.S. Army Iraq and Afghanistan records revealed the phrase “rice bags” – the file analytics tool picked up that IEDs were being hidden under rice bags! This is a case where analytics may have actually saved American lives.

Less dramatically, file analytics would enable the DoN to share information that could be released while protecting what should be withheld. Currently, file analytics on unstructured data is being used by the National Archives for responding to its Freedom of Information Act (FOIA) requests. And for DoN Security and IG offices, this tool would allow for rapid identification of classified and sensitive information, while identifying which information which should be walled off and protected.

---

James Willson-Quayle works in the Office of Strategy and Innovation for the Under Secretary of the Navy (Management)

---

\*\* = The opinions expressed here are solely those of the author, and do not necessarily reflect those of the Department of the Navy, Department of Defense or the United States government.