Big Data over Limited Networks

Lawrence Poynter
Product Management,
iOra Inc,
New York, USA.

Rear Admiral Philip Wilcocks CB DSC DL
Board Advisor,
iOra Limited,
Fleet, Hampshire, UK.

***Abstract*—This paper examines the emergence of Big Data rpositories and the particular problems that are encountered when these data sources need to be transmitted over limited military networks. Following a review of the historical development of Big Data the paper details the technology challenges involved in the management and analysis of large repositories and is illustrated by some existing and emerging military examples including applications in Intelligence, Command and Control, operations, logistics and field maintenance. In an attempt to offer options for dealing with Big Data an evaluation of some of the techniques that can be applied to data to overcome the network constraints that exist are examined.**

I. INTRODUCTION

A. Big Data Development and History

The discussion of large datasets and the potential problems of management, storage and analysis has been running for over 70 years. As early as 1944 Fremont Rider [1] commented that a university library would double every sixteen years, where he speculated that the library in Yale would grow to 200 million volumes in 2040. A set of volumes would require the construction of 6,000 miles of shelving and would need to be managed by a staff of over 6,000 specialists.

The move into the digital age extended our capacity to archive and store information and, with the cost of data storage plummeting, our ability to record digitally has dramatically increased. *In 1980*, Tjomsland [2] described the way in which people had changed the way that they manage data. Tjonsland proposed that people either did not have the tools or time to properly manage their data thus opting to leave obsolete data in place rather than performing a managed pruning process. It would appear that at even at that time that our ability to manage data had been outpaced by the capacity for data to be produced.

Later in 1997 Michael Lesk [3] estimated the total size of storage required for all time. He concluded that by 2000 a few thousand petabytes would be required to record all knowledge that including film, photographs, documents and sound. He concluded that by 2000 the hardware storage market would have matured to the point whereby all data could then be recorded with adequate spare capacity. With the benefit of hindsight it would seem that Lesk was being more than optimistic and the concept of Big Data emerged as an area of study following the presentation by John Masey [4] in 1998. Research continued to estimate the scope and size of the emerging Big Data problem. In 2000 Lyman and Varian [5] published a paper that contradicted the Lesk estimates just three years earlier. In the paper they claimed that the world produced about 1.5 Exabyte's of unique information in the preceding year. The study was repeated in 2003 where the estimate was extended to over 5 Exabytes of information that was now almost entirely digital in nature. According to their research the dominancy of digital as the primary source for storing information was now complete. Their research was later corroborated by the work of Hilbert and Lopez in 2011 [6], who estimated that, in 1986, 99.2% of all storage was analog, but in 2007, 94% of storage was digital.

According to the most current IDC research, the forecast for the global data storage requirement in 2020 is 35 zetabytes where the majority of data will have been generated by the enterprise.

B. Big Data Challenges

From a pure data management perspective big data sets present large scale management problems:

- ***Recording*** – The standard problem of maintaining the storage of larger and larger sets of data where the backend storage infrastructure requires continual upgrade and management to ensure access and backup
- ***Data cleansing*** – Providing data filtering mechanisms to ensure that validity of the data being added: this means that data being added has to be cross referenced with existing data to verify accuracy. The problem is complex with small scale data and becomes considerably more complex as data volumes increase.
- ***Analysis and interpretation*** – Underlying perhaps the primary opportunity for Big Data, the analysis and interpretation of data is challenged by the sheer scale of the data to be processed. Innovative data mining technologies and visualization techniques are used to support knowledge discovery where ever more powerful technologies are required to run the processes.
- ***Archive management*** – The sheer scale of Big Data compounds the task of managing end-to-end lifecycle of data to the point where data is either archived or removed when considered no longer of any value.

- ***Transporting*** – Perhaps the more testing challenge for the military sector is the problem of transporting the data once gathered to a site that can perform either formal storage, analysis, interpretation, or archiving. Military communication networks vary considerably in their capability for transmitting large volumes of data. Typically the networks at the tactical edge are limited in either capacity or availability and hence do not lend themselves to the access of large volumes of data.

C. *Military Big Data and the Opportunities*

The development of significant volumes of data is not restricted to the commercial world: military data repositories can be vast – especially during periods of operational engagement. Furthermore, the requirement to conduct detailed post-combat analysis requires access to vast quantities of data prior to, during and post combat operations. The typical datasets that are generated include the following:

- ***Intelligence*** – The modern military campaign is centered on the collation, distribution and analysis of intelligence data. In modern intelligence-led campaigns, being able to access, process, and action data is a fundamental requirement at all levels from tactical through operational and up to the strategic level. The introduction of modern surveillance systems such as Unmanned Aerial Vehicles (UAV) have created an explosion in the quantity of data available to analyze and process swiftly to maintain operational tempo and momentum. Where UAVs stream video data over a local high speed networks, as much as 90% of the available video content is not viewed at the time of generation thereby missing critical time-sensitive intelligence.

  A fundamental bottleneck in the collation of all this intelligence data is the ability to transmit the data to those individuals who have the skillset to turn intelligence into operational action.

- ***Logistics and Support*** – Military operations involve the rapid deployment of assets to any location on the globe. As a result, the huge pressure on the logistic train requires prioritization of spares and equipment. To maintain high readiness and availability, the overall health status of these assets include embedded sensors that report on status, usage, location, etc. In addition, key support and welfare data critical to operational welfare and ammunition status.

- ***Command and Control*** – The underlying data that is used to compile the Command and Control operational picture is an increasingly complex and data intensive beast. It is one matter to ensure that the information can be readily interpreted and quite another to ensure that all military consumers have equal levels of access irrespective of location or network connection. Immediate benefits can be achieved by ensuring that all users can collaborate and engage on the development of a plan or mission.

- ***Operational data*** – As an interesting example of how the benefits of analysis can now be used to create new insights – weather data recorded back to the days of Admiral Nelson can now be integrated into the Big Data analysis models to support more accurate forecasting. Previously these datasets were excluded for reasons of scale, but now they are making a contribution to military operations 250 years after they were originally put to paper.

II. MILITARY NETWORK ALTERNATIVES

The following section details some of the tactical network options currently available and used by the military for distributed data in the operational theatre.

A. *Cellular*

Military and Government Agencies have been deploying cellular networks for some years with great success. Pre-configured cellular 'green-box' networks that only require the addition of a power source to operate have revolutionized quick and simple wireless communication for deployed forces. Although dependent on the capabilities of antennae that have been erected, once operational these networks can support line of sight communication to a range of 2 to 5 km. In addition these networks typically use wireless communication protocols that support communication to the type of devices most suited to the modern war fighter e.g. smartphones and tablets. Network data capacity can be configured to enable the transport of significant levels of data; however, the limitation on communication range restricts the ability to extend beyond the deployed HQ or FOB. These range limitations are resolved in their sister civilian networks by the installation of repeater stations that effectively bridge the transmission gap between communicating hubs. Repeater stations are not an option for deployed military networks as each station represents an operational risk, requiring physical security for protection in often remote and difficult to support regions. For this primary reason, cellular networks are restricted to provide communications at fixed installations to local patrols that are delivering close operational support.

B. *Radio*

The use of radio based communications has long been, and remains to be, the backbone of deployed military communications. The forms of communication can be broadly grouped into Ultra High Frequency (UHF)/ Very High Frequency (VHF) and High Frequency (HF).

UHF/VHF – High speed communications networks have the ability to transmit at significant data rates. Vulnerable to blocking, these share similar deployment constraints as the cellular networks regarding the requirement for LOS antennae, although they have more potential for connectivity range.

HF – The stalwart of the military community and the dependent bearer for deployed operations. The benefits of HF are significant range e.g. 5000kms but there are limitations in both bandwidth and consistency.

C. *Satellite*

Military communication satellites have increased usage for the last three decades. Most military powers have access to their own sovereign satellite networks, e.g. Skynet5 in the UK or MilStar in the US, although network demand typically results in the use of commercial satellite capabilities e.g. Inmarsat to cater to peak usage.

Typically satellite network capability suffers from regional variations according to the flight pattern of the satellite i.e. Low orbit (limited regional coverage, high bandwidth) or High orbit (significant regional coverage, lower bandwidth). One significant factor in satellite network usage is the cost of operation where the commercial platform is required. It has been estimated that the current global spending on military communications is over $17B per annum, so being able to efficiently manage data over the available channel becomes a key requirement, especially in the increasingly cost conscious focus of recent times.

III. STRATEGIES FOR OVERCOMING NETWORK LIMITATIONS

What this means for those responsible for information management is that they fight their own daily battle to ensure that information is consistent across all deployed sites and command posts with guaranteed availability. This is complicated further where deployment-wide network governance is typically non-existent, meaning that users often have to speculate on the availability of a network to support their specific operational requirement.

The commercial world has evolved to provide some technology solutions to ease difficult networking environments that include the following:

A. *Network Accelerators*

Network accelerators are positioned at either end of a network and have the effect of speeding up communication between two points in that network. In general these devices ‘intelligently’ store repeated network calls issued by the computer so that in effect less data is required to be sent over the network. Most accelerator devices are installed as hardware appliances at each end of the network, although there are some providers that implement a software-only install. These appliances have the effect of typically speeding up network traffic 6 to 10 times. A significant drawback of these devices is that they require a continual network connection to operate and do not proactively forward deploy complete content replicas in the event of network disruption or complete disconnection.

B. *Data Compression*

Reducing the quantity of data that is required to be sent over the network has a direct impact of the bandwidth usage and (commercially) the cost of delivery. Various compression tools are commercially available that provide mechanisms for reducing the data footprint of updates so that better use can be made of the available mobile network. The best compression techniques involve extracting redundant data that does not need to be transmitted from the complete dataset as a whole. This is in comparison with far less effective techniques that analyze delta changes to single files in isolation. The effect of compression can be quite dramatic where files, typically based on Microsoft PowerPoint, are being updated, saved as new instances and then propagated over the network. Where content revisions are limited, the level of file transmission reduction can reduce from 10’s of megabits to the 10’s of kilobits. This in turn has a dramatic effect on the requirements of available network when transmitted onto, for example, 100 forward units.

C. *Content Distribution*

To remove the requirement for the deployed user to reach back over the battlefield network to access data, content distribution is used to proactively deploy key data closer to the user so that they do not need to rely on an external network connection. In this way, for example, an operator would replicate updates to the mission plan on a schedule to the forward operating base, so that, when required by the commander, they have a local store of information and do not have to reach back to access the necessary data. These local data replicas are either based on the same HQ infrastructure or, more increasingly, are implemented using virtualization technology that reduces the administration overhead for both setting up and maintaining the backend portal infrastructure. These virtual platforms can quickly be deployed and re-deployed as and when required.

D. *IP Networks over Radio*

Extending the use of radio to transmit data in addition to voice has become particularly achievable as part of global deployments. However, although presenting a new communication data channel, the bandwidth available is often too restrictive in terms of its ability to adequately provide a network service for a host of web based applications. Pre-processing and post processing in the form of compression and de-compression is required to expand the set of data applications that can be used over the radio network. One significant bonus of radio communication is the perceived cost savings that can be achieved by transmitted data over radio as opposed to costly satellite based networks.

E. *Least Cost Routing*

When a force has access to multiple communication routes, actively switching between providers of bandwidth, using least cost routing is often favored by militaries as a smart way of reducing bandwidth costs and ensuring network availability. This approach has traditionally been used by Navies, where the typical scenario is for vessel communication to switch from satellite based delivery whilst operating offshore blue waters, to more cost effective VHF delivery of the same data when in range of shore (typically 50 to 70 miles). The intention with these programs is that the bearer switch be performed seamlessly by the intelligent bearer hub to ensure both cost savings and network consistency.

F. Hybrid

Typically most militaries use a hybrid of the options described above to ensure that content is consistently available to all interested users are their point of need. A common infrastructure is the use of network acceleration devices alongside deployed replica's, where the acceleration provides optimized access during times of network connection and the deployed replica can either provide LAN speed access at the remote end of the network or fulfill a local replica or COOP (Continuation of Operation) server capability in the event of network disconnection.

IV. CONCLUSIONS

Military operations will need to continue to manage ever larger repositories of data and information. Exploitation of the benefits of increased data analysis to deliver operational effect will become an ever increasing military objective, in order to achieve military supremacy. The bottleneck of existing tactical networks will need to be addressed in the battle for transmitting the data to where it can be analyzed and interpreted. Significant research and development investment is being expended in all forms of wireless networks to support remote military deployments. The transmission of data both up and down the battlefield chain of deployed command will continue to be a critical requirement for armed forces that operate globally. In parallel, the ability for data production at all nodes of the network will increase the burden on the deployed network and we will continue to require techniques for ensuring that access to key data is provided. This means, whatever the location, information consistency and the single point of truth are ensured.

REFERENCES

[1] F. Rider, "The Scholar and the Future of the Research Library." Bulletin of the Medical Library Association January 1945.

[2] I.A. Tjomsland "Where Do We Go From Here?" Fourth IEEE Symposium on Mass Storage Systems, April 1980

[3] M. Lesk publishes "How much information is there in the world?" www.lesk.com, 1997

[4] J. R. Masey, "Big Data… and the Next Wave of Infrastress." USENIX Meeting, April 1998

[5] P. Lyman, H.R. Varian "How Much Information?" UC Berkeley October 2000

[6] M. Hilbert, P. L. publish "The World's Technological Capacity to Store, Communicate, and Compute Information" *Science*. February 2011

[7] IDC "Digital Universe Study" September 2011