Lessons Learned from Initial Exploitation of Big Data and AI to Support NATO Decision Making

Michael Street, Peter Lenk, Ivana Ilic Mestric
NATO Communications and Information Agency (NCIA), Data Analytics and Innovation
Oude Waalsdorperweg 61
2597 AG The Hague, NETHERLANDS
Michael.Street@ncia.nato.int; Peter.Lenk@ncia.nato.int; Ivana.IlicMestric@ncia.nato.int

Marc Richter
European Union Agency for Law Enforcement Cooperation (Europol),
Information Architecture[1]
Eisenhowerlaan 73
2517 KK The Hague, NETHERLANDS
Marc.Richter@europol.europa.eu

ABSTRACT

Analysis of NATO data with analytics and machine learning has the potential to improve decision making for NATO commanders. This paper addresses an emerging architecture and lessons learned to analyze NATO data with the '4V' attributes of: volume, variety, veracity and velocity. The focus is to describe the engineering challenges of a capability for data analysis and artificial intelligence when operating in a classified environment. The paper captures lessons learned in establishing such an environment and describing several use cases where analysis has been successfully conducted.

1.0 INTRODUCTION

Exploitation of data analytics in the NCI Agency started with the creation of the NCI Agency, as data analytics tools were used to consolidate and assess data from business applications across the Agency's predecessors. From these initial steps, NCIA's data science and data analytics capability has grown. This paper describes some use cases and architectural implications of the NCIA data science and analytics capability.

Initially the Agency did not analyse "big" quantities of data, but used tools and methods drawn from the big data world. However, a data analytics team within NATO's Communication and *Information* Agency was inevitably drawn to new sources of data and the team had soon established a data mart, fed from many sources, creating a data lake to allow "big" data analysis.

Soon after, a number of NATO users with challenging problems were keen to investigate whether big data or artificial intelligence (AI) could offer a way to solve operational challenges or to extract more value from existing information. In applying big data techniques and AI / machine learning to a variety of NATO problems, a number of lessons have been learnt of how to improve the exploitation of these technologies in NATO.

[1] Formerly with NCIA

2.0 USE CASES

This section gives an overview of use cases where big data approaches and machine learning have been applied to NATO data. This data originates in a range of security classifications; public information, low classification *business* data, and classified data from operational systems.

2.1 Enterprise business intelligence

Business Intelligence (BI) is an umbrella term to describe data analysis and prediction (which may include AI) to improve and optimize business decisions and performance [1]. Early NCI Agency management dashboards drew on data extracted from a range of legacy management information systems used within the five predecessor entities under a project to provide metrics for Performance, Measurement, Analytics, Reporting and Benchmarking (PMARB). The variety of systems, and data types required to consolidate Agency dashboards led to use of data analytics and visualisation tools.

Data sources are typically conventional business applications such as those in support of e.g. finance, project management, HR etc; where high system resilience is required. To ensure data integrity within systems acting as data sources, analysis is not conducted on live system data; instead data is copied from system backups to a data staging area (or replication layer), an operation which normally takes place overnight after daily system backups are made. This prevents any corruption of system data in production environment. From the staging area data is extracted from the database backups, transformed and loaded into the “landing zone” record-level change history tables, which are then used to populate enterprise data marts in the Restricted domain. Data passing through this extraction, transformation and loading (ETL) process is of high volume, large variety and (upon entering the data marts at the end of the process) high veracity. The driver is to extract and visualise Agency-wide insight from this varied dataset.

2.2 Deduplication of large data sets

NATO’s ISAF mission ran for 13 years, during which a wide variety of information systems and functional services created a wealth of data. With the completion of the mission, this data – several hundred terabytes - is being collected and archived. During a long running operation spanning a large geographical area, files are easily duplicated at multiple data storage locations, by multiple users, during frequent backups of both scheduled and unscheduled nature. This results in a significant quantity of information which is duplicated across different locations, organisations and periods. Deduplication of this data is a significant step which dramatically reduces the amount of effort needed in later stages; for long term archive storage, and (crucially) for analysis and tagging with metadata so that information within the archive can be identified when needed. Manual deduplication is a highly repetitive process subject to human error. A demonstration on a subset of data has shown the potential to automate deduplication both on the metadata and the information content levels. The agency analytics toolset has also been used to read functional service archives directly, negating the need to install legacy versions of functional services to access that data.

2.3 Information environment assessment

NCIA activity to support the information environment assessment (IEA) takes the data analytics and visualisation capabilities developed for enterprise business intelligence and adds them to the open information sources provided through the Alliance open source system (AOSS). Open source tools for data analysis are being deployed in the AOSS environment, providing additional functionality to analyse the large quantities of information being streamed into AOSS, which in turn can aid the analysts and other AOSS users. This work is documented in more depth in [2].

2.4 Classification of objects and documents

The ability of data analytics tools, and in particular of machine learning and deep learning to predict information or relationships has been applied in several use cases where NATO information does not have the full complement of metadata.

2.4.1 Security classification of documents

A number of large document sets and information objects exist which have never received a security classification. Data analytics and machine learning have been applied to this problem to explore the potential to predict security classifications for documents based on their content. Open source deep learning models have been applied to a test document set, as have proprietary tools and models. This work is described in detail in [3] but the implications of comparative performance of different models is addressed in section 3.1.

2.4.2 Entity extraction of battlespace objects

Machine learning has also been used for entity extraction of battlespace objects from documents/reports produced during a NATO mission. As a training set we used a subset of classified and processed reports with corresponding battlespace objects produced by analysts. A *supervised learning* algorithm was trained from this data set before being applied to other documents/reports. Figure 1 shows the process implemented in KNIME Analytics Platform [4] to extract and classify the battlespace object using a Random Forest machine learning model.

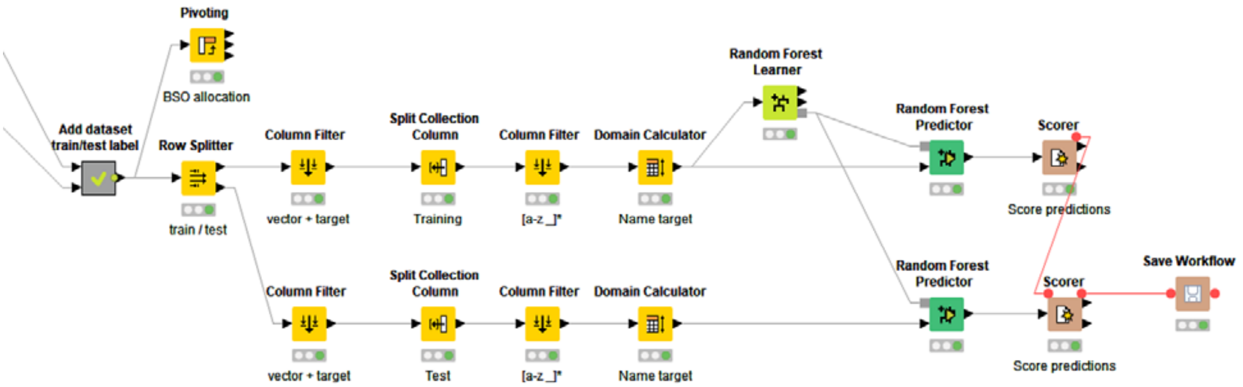

Figure 1: Process flow for automated entity extraction.

2.5 Anomaly detection

Identification of anomalistic behaviour is another area where data analytics and visualisation tools have been used to explore the contents of system log files to extract additional value from this data. These files capture actions of users and administrators of functional services over long periods, logging actions and allowing a rich source of data to explore human behaviours with the system.

Figure 2 shows access patterns and actions in one functional service, grouped by user type and action over a 12 month period. It can be seen that certain user groups have comparable behaviours e.g. second left and right-hand columns. Horizontal lines running through almost all columns show that at certain periods, almost all types of users act simultaneously.

Building up an understanding of regular and predictable user behaviour through analysis and visualisation of logfile data helps identify anomalistic behaviour which could be an indicator of insider threats, other malicious activity or non-malicious actions which require further investigation. It could also be used to generate requirements for updates or future systems.

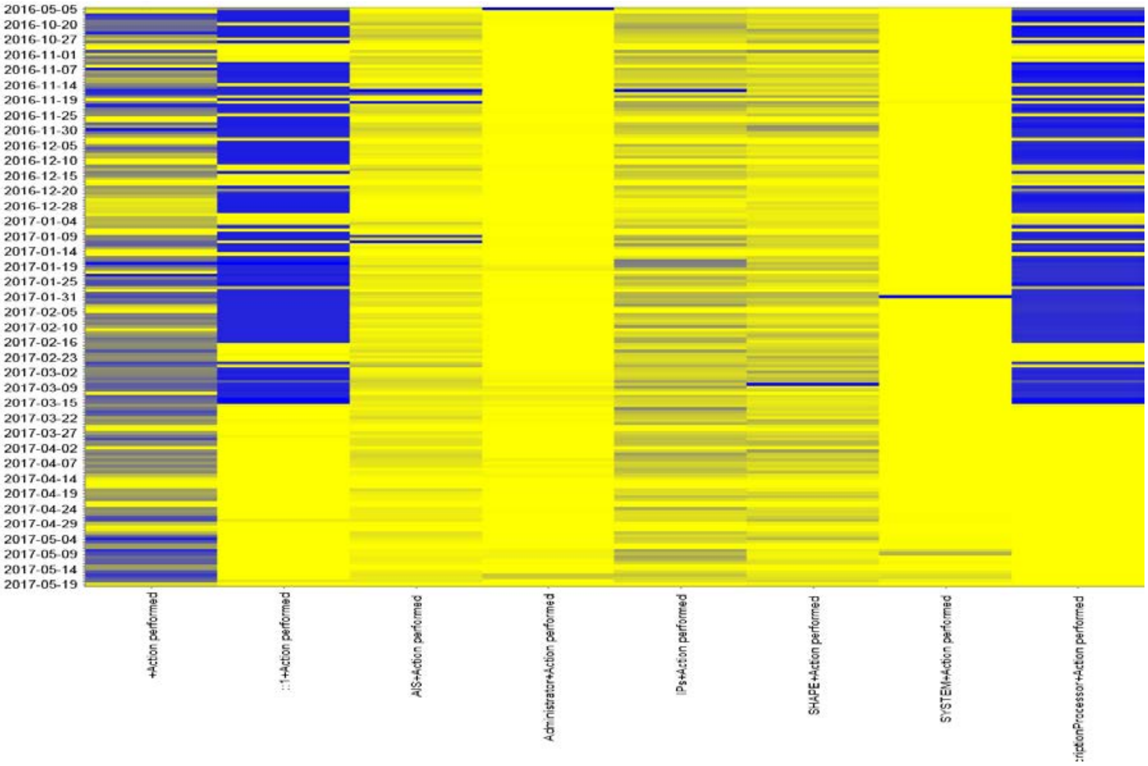

Figure 2: Initial visualisation of logfile activity through time.

2.6 Document comparison

NCIA’s initial application of data analytics to assess the correlation of NATO documents is described in [5]. Since this work, document analysis and correlation work has continued, including work to analyse NATO’s C3 strategy and C3 policies. Analysis of this textual data is used to discover coherence, overlaps and gaps. The current analysis uses text similarity – the extraction of structured knowledge and information from text and analysis of similarities. Text similarity is a metric defined over a set of terms or documents, in this case between NATO strategy and policies, where the distance between them is based on the syntactical representation. Again, the open-source KNIME Analytics Platform tool is used to perform the text mining and analysis. The basic process is shown in figure 3.

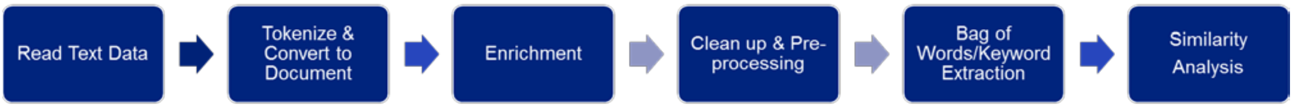
Read Text Data
Tokenize & Convert to Document
Enrichment
Clean up & Pre-processing
Bag of Words/Keyword Extraction
Similarity Analysis

Figure 3: Text mining with KNIME.

Visualising the resulting Bag of Words as a word cloud gives an initial indication of the scope and correlation of the documents. Figure 4 shows the correlation of terms between the strategy and policies documents Microsoft Power BI.

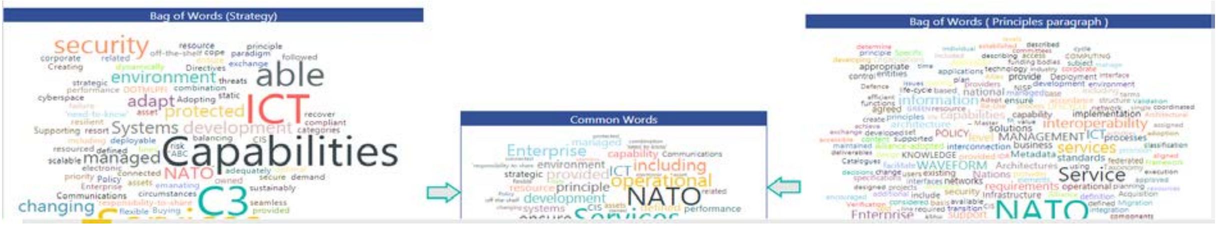

Figure 4: A visualisation of document correlation.

3.0 ARCHITECTURE

Big data is often characterised by the *four Vs* of volume, velocity, variety and veracity. Each of the use cases listed above exhibits at least one of the data characteristics of volume (being too big to handle in any way with conventional systems); velocity (too big to process in a timely manner given the speed of data generation); variety (too big to handle with appropriate distinction) and veracity (too big to handle well and to process correctly by hand). The architecture to address analysis of data with these four characteristics in the NATO restricted domain is shown in figure 5.

3.1 Toolsets and technologies

The toolset used in the NCIA has evolved to focus on KNIME Analytics Platform, an open source data processing tool and Microsoft Power BI for visualising data. Both tools were market leaders in recent Gartner magic quadrant assessments. KNIME Analytics Platform is an open source tool with a powerful graphical user interface, reducing the need for coding by data scientists and analysts; it also provides native support for Python 3.0 which allows integration of deep learning libraries such as TensorFlow [6] and Keras [7] to extend the functionality and resources available.

The security classification use case employed a number of Deep learning and Meta learning libraries e.g. Random Forest (K-RF), Multi-Layer Perceptron (K-MLP), Tree-based Gradient Boosted Method (K-GBM) and Boosted Lasso Logistic Regression (W-LogB) models. Of these open source models, Random Forest provided security classification prediction performance comparable to earlier work on the same data set using a patented Helmholtz algorithm [3]. This shows that there is a role for specialist, proprietary models and tools, but these are not always necessary if a wide range of open source options are available in the hands of sufficiently experienced data science teams.

Figure 5 shows two tools for visualisation, as the Agency recently migrated earlier work using VisionWaves to Microsoft Power BI. This was largely to reduce cost as Power BI was able to replicate the previous dashboards at a fraction of the cost; although the effort to migrate skills and dashboards between the tools was non-negligible. Separation of data preparation from visualisation allowed the visualisation tool to be changed without it impacting data preparation or data sets which significantly reduced the effort to switch tools.

3.2 Volume

Although the capability in figure 5 is designed for data analytics, resources are limited and brute force approaches to data analysis are not possible. So it is important to select appropriate algorithms and models based on a data scientist’s experience and understanding of the data to optimise use of the available resources. This optimisation is an ongoing role as the relationship between the quantity of data and analytic resources is not linear. In cases such as business intelligence the amount of data grows linearly with time; but processing needs (if comparing the latest records to all previous records) will grow exponentially.

Even with toolsets designed for big data applications, complex functions may generate the same result from different inputs due to the huge range of inputs, e.g. where hash functions are used to quickly compare records (for deduplication) it is inevitable that different records in huge data sets will occasionally generate the same hash value. Therefore when comparing records in large data sets matching hash values must trigger a complete check of field values before declaring a match.

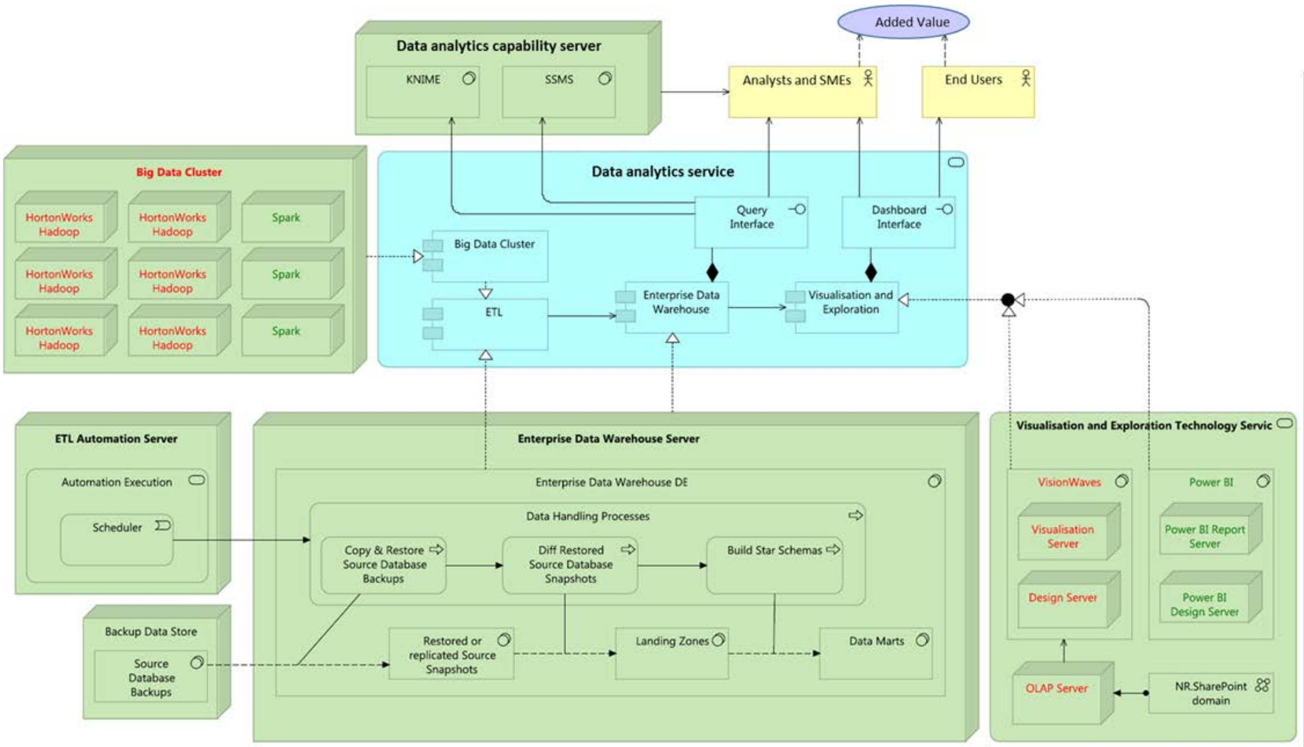

Figure 5: Architecture.

The computationally intensive processes to learn and refine machine learning models used to predict security classification or battlespace objects often make it necessary to truncate the data sets used for training the model, and to restrict the number of models and iterations used for training. This necessarily leads to sub-optimal results from the machine learning process and is discussed further in [3].

3.3 Velocity

Some use cases draw on data which is updated on at least a daily basis, requiring data to be extracted, transformed and loaded (ETL) regularly. Automation of the ETL process reduces the workload on those supporting the data analysis team; hence the inclusion of an ETL automation server (figure 5) to automate this as much as possible.

It is essential for those responsible for ETL to have a close relationship with those responsible for the systems generating data. For example, proactive notification of system updates or modifications which affect the format of data stores or archives will prevent inaccurate data being fed into the analytics, producing inaccurate results in visualisation dashboards.

Experimental deployments of ETL processes have utilised streaming data capabilities, ensuring that further increases in data processing velocity requirements can be addressed with “live” data manipulation capabilities.

3.4 Variety

A variety of data sources allows more accurate and more interesting relationships to be discovered and more questions to be answered with greater accuracy. Many tools exist which allow structured and unstructured data sets to be transformed and loaded into a data mart / data lake. The data scientists responsible for this need input from domain experts on the particular data set to prevent data being misrepresented e.g. systems may use the same field names for different variables. An additional challenge in a NATO environment is the variety of security domains from which data can be sourced. Transferring huge data sets across security domains (only ever to a higher domain) poses unique technical and procedural challenges.

3.5 Veracity

Accuracy of source data is a continual challenge. Technical factors either in the source systems or in an overly complex ETL process can impact data quality, as does the commitment of system users to maintaining their data. For some functional services the data within the system may be subjective with different analysts describing the same object in different ways; this presents a challenge to data analysis and especially to machine learning. Several use cases show analytics and machine learning can be applied to existing data to identify differences in classification by humans to both to help increase data quality and also to identify improvement areas for the system or for user training.

4.0 EXPLOITATION OF BIG DATA & AI TO SUPPORT DECISION MAKING

In addition to the *four* Vs referred to above, a fifth *V* of *value* is increasingly referred to. Data science and technologies should be able to provide value for decision makers, whether they be in operations, securing information or guiding NATO activities. The value provided may depend on whether the technology replaces, augments or supports humans in complex tasks.

Work by the NCI Agency in recent years has demonstrated that there is no single tool which can solve all NATO’s big data and AI problems; but a small suite of cost-effective tools which support effective integration can provide a powerful toolset.

Any data science technology is dependent on data. Making NATO data available beyond the source system at scale is a new concept. This will become a future role of system managers, while future systems will need to provide *analytics ready* data to minimise the ETL effort required, allowing data scientists to maximise value from analytics and AI, rather than rectifying the problems of data generation by legacy systems.

Section three noted the need for data scientists to ensure that data is sourced and processed accurately, and to work with system owners to increase the veracity of data as this will increase the value of the output. Equally, decision makers and other end users of data analytics and machine learning must be aware of the caveats on the effective use of these technologies.

The correlation between machine learning and human assessment is rarely 100%, correlation over 70% is considered very good in many use cases (although comparable correlation between humans is rarely assessed). Such limited correlation makes machine learning unlikely to fully replace humans in decision making. Instead the value of machine learning in these scenarios is not to make the decision for a human, but to analyse and indicate areas of interest, providing an overview to a decision maker and indicate where to focus attention. In these scenarios machine learning is providing decision support, but not decision making.

ACKNOWLEDGEMENTS

This paper and the work within it would not be possible without a great many colleagues. Without past and present members of the NCIA data analytics team the Agency could not have addressed the challenges described: many thanks to Arvid Kok and Giavid Valiyev. Nor would we be learning lessons of data analysis without data to analyse and the domain expertise to explain and explore it; thank you ACO J2, Kurt Veum, Lee Houlker, Phil Eles, Konrad Wrona, Birger Retzius, Eleanor Williams and Marcello Piraino.

REFERENCES

[1] Gartner, https://www.gartner.com/it-glossary/business-intelligence-bi/, 2018

[2] R. Blunt, C. Riley, M. Richter, M. Street, D. Drabkin, “Using data analytics and machine learning to assess NATO’s information environment”, IST-160 specialists meeting on *Big data and artificial intelligence for military decision making*, Bordeaux, May 2018.

[3] M. Richter, M. Street & P. Lenk, “Deep Learning NATO document labels: a preliminary investigation”, Int. Conf. on Military CIS, Warsaw, May 2018.

[4] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kotter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, "KNIME: The Konstanz Information Miner," in Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 2007

[5] P.T. Eles, B. Pennell, M. Richter, “Assessing NATO policy alignment through text analysis”, Int. Conf. on Military CIS, Brussels, May 2016.

[6] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, ‘TensorFlow: A System for Large-Scale Machine Learning” in OSDI, vol. 16, pp. 265-283, 2016.

[7] F. Chollet, “Keras: Deep learning library for theano and tensorflow,” https://keras.io, 2015.