ENTERPRISE DATA ANALYTICS

HO Wee Peng Ricky, KOH Lay Tin, CHONG Yonghui Benjamin, HO Chi Ming

ABSTRACT

Organisations today have an increasing desire to derive insights from their data in order to make better, real-time, data-driven decisions. However, even with the increasing data availability and technological advancement, many organisations are still unable to realise the full potential of Data Analytics (DA). This is especially so for large organisations, where the enterprise architecture landscape is usually highly complex and comprises fragmented systems designed to optimise transactions within data silos. These inhibit the cohesive collation and dissemination of data, which is a key success factor for DA. In addition, the emergence of a myriad of DA technologies and tools exacerbates the difficulty in implementing DA capabilities in a coordinated and sustainable manner. Much effort has also been invested into the development of analytics model to find new insights, but many neglect the need to have a systematic approach to operationalise the capabilities. This article describes how DSTA identified and put in place the essential data, technology and people enablers to ensure a continuous stream of sustainable data analytics capabilities.

Keywords: data analytics, big data, data lake, data governance, data scientists, data analysis

INTRODUCTION

DSTA has developed Data Analytics (DA) capabilities for the Ministry of Defence (MINDEF) and the Singapore Armed Forces (SAF) to support the administrative functions for resource management, policy formulation, governance and decision making. The emergence of new technologies, increasing wealth of data, and the development of data analysis techniques have given rise to opportunities to seek new and innovative use of data. A robust, scalable and sustainable data and analytics architecture is essential to enable the development and delivery of new capabilities.

ENABLERS FOR SUSTAINABLE DA

There are three key enablers (see Figure 1) for DA development: technology, data, and people. This paper discusses how DSTA strengthened the strategic enablers for the coherent development of new DA capabilities in MINDEF.

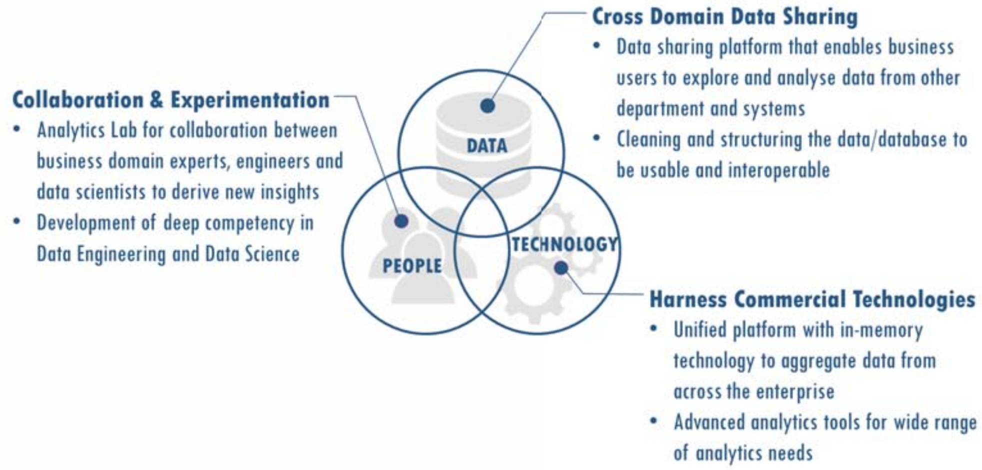

Cross Domain Data Sharing
• Data sharing platform that enables business users to explore and analyse data from other department and systems
• Cleaning and structuring the data/database to be usable and interoperable
Collaboration & Experimentation
• Analytics Lab for collaboration between business domain experts, engineers and data scientists to derive new insights
• Development of deep competency in Data Engineering and Data Science
DATA
PEOPLE
TECHNOLOGY
Harness Commercial Technologies
• Unified platform with in-memory technology to aggregate data from across the enterprise
• Advanced analytics tools for wide range of analytics needs

Figure 1. Enablers for DA

ENTERPRISE DATA ANALYTICS PLATFORM

The Enterprise Data Analytics Platform (EDAP) as shown in Figure 2 is a comprehensive solution designed and implemented by DSTA Enterprise IT Programme Centre to put in place technology, data and people enablers for the development of DA capabilities in a coordinated and sustainable manner.

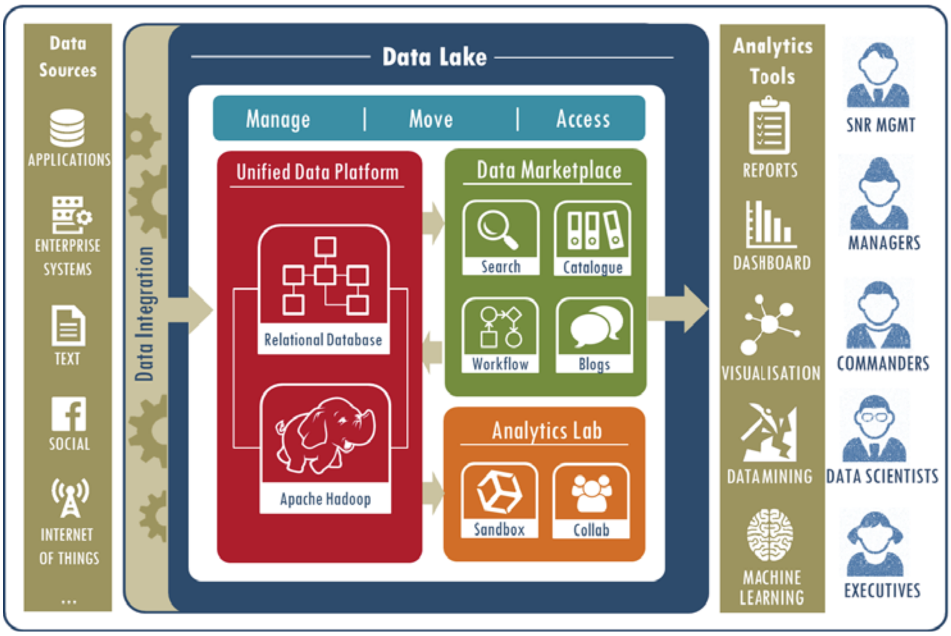


Figure 2. EDAP

TECHNOLOGY ENABLERS

Unified Data Platform

Enabling DA at the enterprise level requires the recognition that different line of businesses (LOB) and MINDEF Corporate IT (CIT) users have varying data requirements. The solution will need to enable data scientists to process big data sets on the fly, glean strategic insights, and also continue to provide structured data marts for corporate users to answer their questions. The challenge was for the EDAP team to deliver an evolvable data architecture that is scalable, efficient, cost-effective and one that can meet the requirements.

Hybrid Design

The team adopted a hybrid approach in the form of a unified data platform which comprises a high performance in-memory database, and an engineered Apache Hadoop platform for big data requirements. They were integrated via a high speed network fabric to allow seamless data queries to be made on data sets distributed across the two platforms. The solution's modular design also allowed the team to scale it easily for future demands by adding computing and/or storage nodes.

Mix Workload Solution

One of the design considerations for the EDAP was the reduction of downstream effort to integrate data generated by transactional systems. The decision was to design the unified data store to support mix database workload, instead of deploying a purposeful "edge" system for analytics only. With this approach, the EDAP team was able to consolidate over 17 database servers supporting over 40 applications in MINDEF CIT into two unified data platforms in the MINDEF and Logistic Enterprise data centres. The ability to ring-fence transaction data as they are generated reduces the resources and effort needed for system management and data integration to meet DA requirements.

Data Archival

Another key principle that guided the design of the EDAP was the need to consider all data as being useful for analysis, for now or in the future. The unified data platform enabled the archival of historical data from the database platform to the more cost-effective Apache Hadoop platform. This enriched the data lake for data scientists to perform deeper analysis, discover unknown trends and find answers to longer-term strategic problems.

Enterprise Analytics Tools

The EDAP provides a suite of enterprise analytics tools that is tightly integrated for descriptive, diagnostic and predictive analytics. This is to meet the end-to-end DA requirements for different groups of users, ranging from information consumers, power users, to data analysts and scientists.

Dashboard and Reporting Tool

The dashboard and reporting tool allows data analysts to create charts and summary tables in dashboards and reports using clean and trusted data that is constantly updated in the back end. These standardised dashboards and reports are used to monitor metrics, as well as Key Performance Indicators for operations and decision making.

Data Discovery Tool

The data discovery tool allows data scientists, analysts and end users to carry out self-service data exploration and visualisation. The tool's intuitive interface and its in-memory capabilities enable the quick analysis of data from many different perspectives through searching, slicing and dicing of data in real time. In addition, back-end data can also be blended with personal flat files or Excel files to perform deeper data analysis and exploration.

Statistical and Predictive Analytics Tools

The statistical and predictive analytics suite of tools enables data analysts and scientists to carry out data cleansing and transformation, apply descriptive and inferential statistical techniques, and employ supervised and unsupervised machine learning techniques to classify data, discover patterns and predict future outcomes. These tools provide superior performance and scalability with multi-threaded operations on large amounts of data, and are extensible with integration support for open-source languages such as R and Python for niche requirements.

Content Analytics Tools

Increasingly, many DA projects use textual information for analyses to derive insights. The EDAP team recognised this trend and incorporated a suite of content analytics tools that allows data analysts and scientists to apply machine learning, text mining and natural language processing on textual data to identify key words, concepts and sentiment.

Big DA Platform

The big DA set of tools allows data scientists to apply statistical, machine learning, text analytics techniques on large data sets (over two gigabytes) stored in the big data platform by leveraging its distributed and parallel processing capabilities. The platform also provides stream processing capabilities for Internet of Things data sources.

DATA ENABLERS

Cross Domain Data Sharing

DSTA has been developing and deploying enterprise systems across the 13 LOBs in MINDEF CIT to automate administrative processes and resource management. The rich source of structured data provides the foundation for DA. The MINDEF Data Store (see Figure 3) enables data sharing across business domains to derive new insights. For instance, using training and medical data can allow one to have a better perspective of our national servicemen's injury trend and its causal factors.

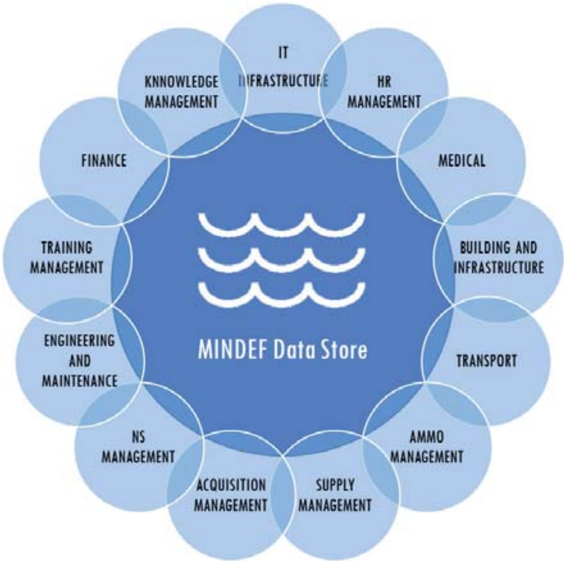


Figure 3. Data lake for 13 LOBs

Data Governance by Design

To enable interoperability of data across the LOBs, the EDAP adopts a data governance by design approach to ensure all data can be integrated for analysis. All data coming on board the EDAP are required to conform to the defined data standards, based on the master data definition that is governed by the data managers from the respective domains. The process ensures all data flowing into the data lake adheres to data governance policies, and eliminates any downstream data ambiguity.

MINDEF Data Store Portal

To enable secure data sharing, the EDAP team designed and developed a portal through which users could come to source for data from a catalogue, request for them via a workflow, and work on them with analytics tools upon approval from the data owners. All these were made available through a single web portal that is integrated with the data and analytics infrastructure. The features of the data store portal are as follows:

Improved Visibility

The portal is integrated with the metadata repository to publish information about the data that is captured in the unified data store. Integrated search capabilities provide users with a seamless process in their "data shopping" experience. A shopping cart concept was adopted to improve its usability from browsing, to requesting and seeking approval for access to data set(s).

End-to-End Automation

An integrated workflow (See Figure 4) was designed to link up the data requestor, supervisor and data owner for the data request and approval process. This reduced the time taken for data sourcing, request and approval. The approved data sets would be automatically provisioned from the unified data store to the enterprise DA tools without the need for users to download the data onto their computers.

Integrated Security

The data shared on the portal is stored in the unified data platform and governed according to MINDEF's security requirements, unlike public sharing portals where data is made available in flat files for download. Data requestors are required to indicate the need and period of use for the requested data, and the system tracks and revokes expired subscriptions. Access to the data is also revoked upon staff resignation or change in appointment.

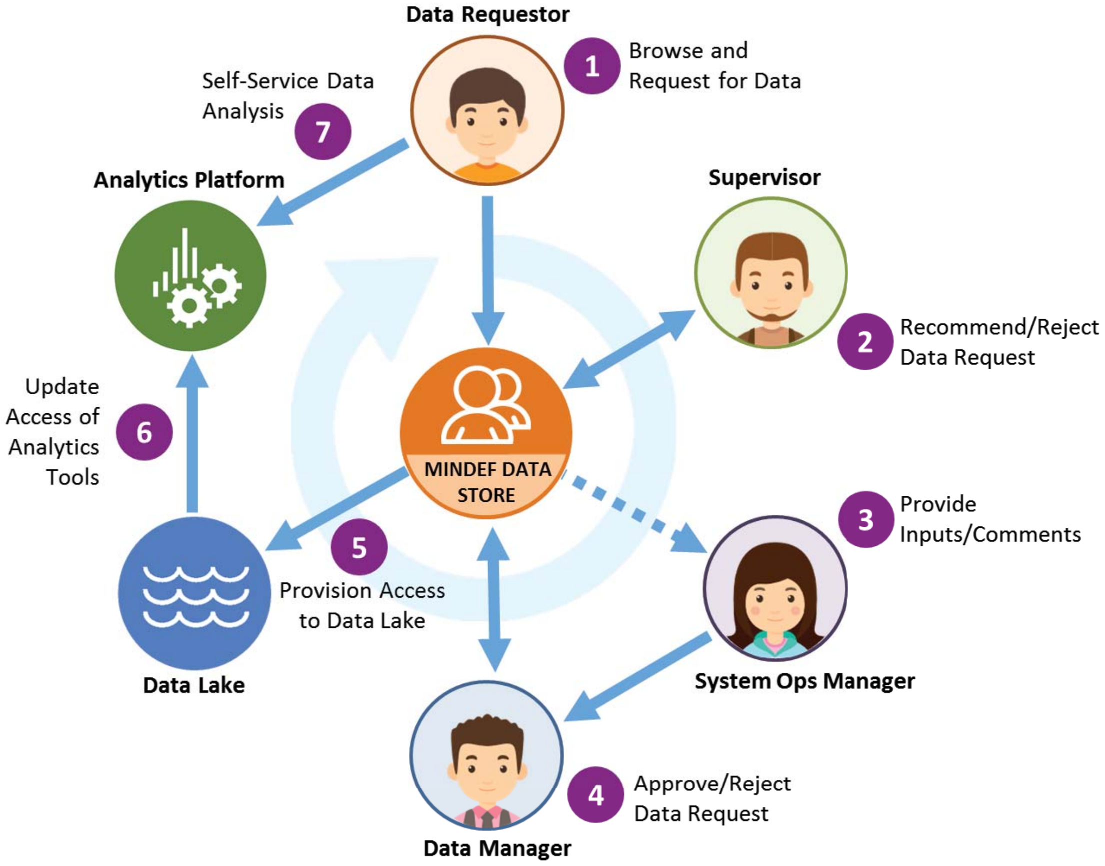

Data Requestor
1
Browse and Request for Data
Self-Service Data Analysis
7
Analytics Platform
Supervisor
2
Recommend/Reject Data Request
MINDEF DATA STORE
Update Access of Analytics Tools
6
3
Provide Inputs/Comments
5
Provision Access to Data Lake
Data Lake
System Ops Manager
4
Approve/Reject Data Request
Data Manager

Figure 4. Workflow for the MINDEF data store portal

PEOPLE ENABLERS

Analytics Lab

DA is a largely iterative process that requires close collaboration between stakeholders, domain experts, engineers and data scientists. The EDAP provides a conducive environment that allows the DA community to thrive through an analytics lab which comprises a sandbox environment and a collaborative space. An artist's impression of the working area of the Analytics, Collaboration and Experimentation (ACE) Lab is shown in Figure 5.

Figure 5. Working area of ACE Lab

Sandbox Environment

The sandbox environment is equipped with analytics tools, computational power and security that are similar to the EDAP's offerings, and loaded with desensitised production data for experimentation. This logical separation of computing resources creates a dedicated environment for data scientists to explore real data and experiment with new technologies without interfering with mission-critical operations.

Collaborative Workspace

The collaborative space (see Figure 6) in the lab is equipped with configurable furnishing, interactive projection equipment and other amenities to enable scientists, business owners and domain experts to collaborate easily without operational concerns. The set-up encourages agile development, improves communications and promotes innovation.

Figure 6. Collaborative space in ACE Lab

Technical Competencies

DSTA recognises the need to develop technical competencies in defence technologies through the creation of the competency development framework in 2003. The framework was updated in 2017 to include the Data Science and Data Engineering domains, reflecting the need and commitment to develop deeper DA competencies to support requirements in DSTA, MINDEF and the SAF.

DA Training

A DA Training framework was also developed to equip different groups of staff in DSTA with the necessary skills for effective application of DA in their area of work. A total of 220 staff were trained in Financial Year 2017, with the objective of equipping all DSTA staff with a fundamental understanding of DA.

DA CAPABILITY DEVELOPMENT

The three enablers provide the necessary ingredients for DA capability development, namely: a data foundation with a wealth of data describing the organisation's business and operations, a suite of analytics tools from data processing to event prediction and insights representation, and a pool of trained DA staff in a conducive working environment. This strategy has enabled the development of new capabilities for the Fleet Management System (FMS) and the Next Generation Procurement System (NGPS).

Leveraging DA in FMS

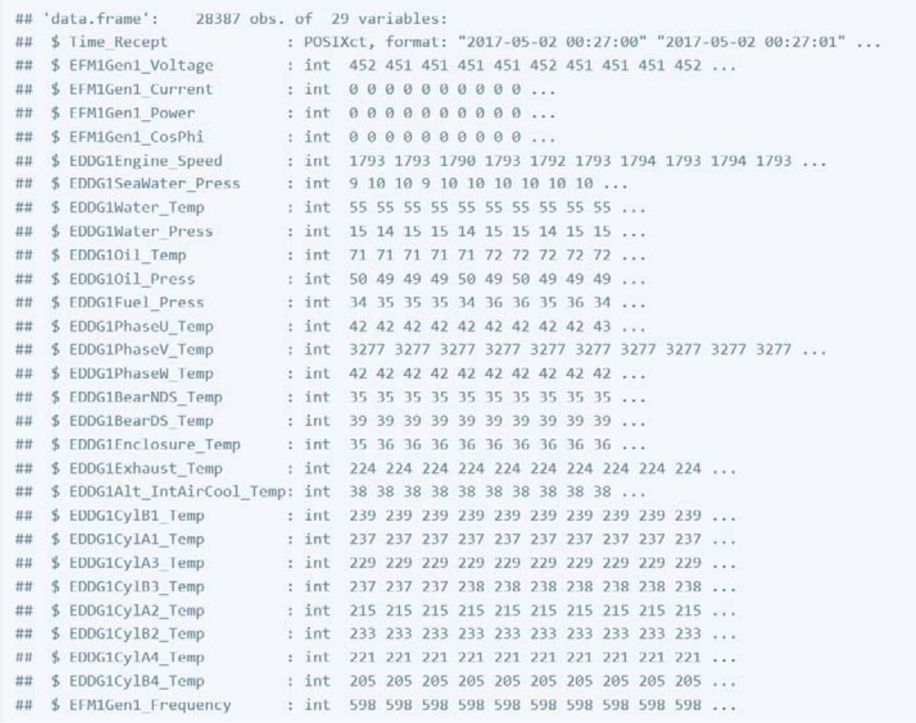
The FMS is a DSTA initiative to tap the increased availability of system, environmental, logistics, engineering and administrative data to improve maintenance efficiency and operational readiness. The following is a case study on the use of DA to detect anomalous behaviour in the diesel generator.

Data Collection and Preparation

Telemetric data from a diesel generator on board a navy platform was collected on several dates where the vessel was in operation. The data described an oil leak due to a cracked oil pump, and the observations made after rectification. The format of the data collected, as shown in Figure 7, was loaded into the sandbox environment for additional data preparation and further analysis.

Applying Machine Learning Techniques

The Support Vector Machines (SVM) is a supervised machine learning algorithm for computing optimal decision boundaries that separate data points for decision making. The team was able to use this technique to derive the linear decision boundary and zone as shown in Figure 9. Readings of the oil pressure and temperature in the green zone represent normal behaviour of the diesel generator and readings in the grey zone represent anomalous behaviour.



```
## 'data.frame':    28387 obs. of  29 variables:
##  $ Time_Recept            : POSIXct, format: "2017-05-02 00:27:00" "2017-05-02 00:27:01" ...
##  $ EFM1Gen1_Voltage       : int  452 451 451 451 451 452 451 451 451 452 ...
##  $ EFM1Gen1_Current       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ EFM1Gen1_Power         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ EFM1Gen1_CosPhi        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ EDDG1Engine_Speed      : int  1793 1793 1790 1793 1792 1793 1794 1793 1794 1793 ...
##  $ EDDG1SeaWater_Press    : int  9 10 10 9 10 10 10 10 10 10 ...
##  $ EDDG1Water_Temp        : int  55 55 55 55 55 55 55 55 55 55 ...
##  $ EDDG1Water_Press       : int  15 14 15 15 14 15 15 14 15 15 ...
##  $ EDDG1Oil_Temp          : int  71 71 71 71 71 72 72 72 72 72 ...
##  $ EDDG1Oil_Press         : int  50 49 49 49 50 49 50 49 49 49 ...
##  $ EDDG1Fuel_Press        : int  34 35 35 35 34 36 36 35 36 34 ...
##  $ EDDG1PhaseU_Temp       : int  42 42 42 42 42 42 42 42 42 43 ...
##  $ EDDG1PhaseV_Temp       : int  3277 3277 3277 3277 3277 3277 3277 3277 3277 3277 ...
##  $ EDDG1PhaseW_Temp       : int  42 42 42 42 42 42 42 42 42 42 ...
##  $ EDDG1BearNDS_Temp      : int  35 35 35 35 35 35 35 35 35 35 ...
##  $ EDDG1BearDS_Temp       : int  39 39 39 39 39 39 39 39 39 39 ...
##  $ EDDG1Enclosure_Temp    : int  35 36 36 36 36 36 36 36 36 36 ...
##  $ EDDG1Exhaust_Temp      : int  224 224 224 224 224 224 224 224 224 224 ...
##  $ EDDG1Alt_IntAirCool_Temp: int  38 38 38 38 38 38 38 38 38 38 ...
##  $ EDDG1CylB1_Temp        : int  239 239 239 239 239 239 239 239 239 239 ...
##  $ EDDG1CylA1_Temp        : int  237 237 237 237 237 237 237 237 237 237 ...
##  $ EDDG1CylA3_Temp        : int  229 229 229 229 229 229 229 229 229 229 ...
##  $ EDDG1CylB3_Temp        : int  237 237 237 238 238 238 238 238 238 238 ...
##  $ EDDG1CylA2_Temp        : int  215 215 215 215 215 215 215 215 215 215 ...
##  $ EDDG1CylB2_Temp        : int  233 233 233 233 233 233 233 233 233 233 ...
##  $ EDDG1CylA4_Temp        : int  221 221 221 221 221 221 221 221 221 221 ...
##  $ EDDG1CylB4_Temp        : int  205 205 205 205 205 205 205 205 205 205 ...
##  $ EFM1Gen1_Frequency     : int  598 598 598 598 598 598 598 598 598 598 ...
```

Figure 7. Telemetric data collected from the diesel engine

Exploratory Data Analysis

After data preparation, an exploratory data analysis was done using visualisation tools. The process enabled data scientists to better understand how the oil pressure and temperature behaved on the days the leak persisted and after rectification.

The analysis in Figure 8 shows distinct patterns in the oil temperature and oil pressure parameters when the oil leak was detected on 2 May 2017 (represented by red dots), when the oil leak persisted on 4 May 2017 (represented by green dots), and when it was rectified on 19 May 2017 (represented by blue dots). To enable the development of a predictive model to detect the behaviour of the engine oil leak, an optimal boundary was defined to separate normal oil pressure and temperature readings from the anomalous ones.

The data was divided into a training data set to develop the predictive model, and a testing data set to validate the results. As shown in Figure 10, the developed model was able to predict accurately the date where oil readings were abnormal on 8 May 2017 (represented by red dots), and the date when the engine had been rectified and readings were normal on 21 May 2017 (represented by blue dots). The analysis proved that machine learning techniques can be applied to detect oil leakage on board the navy vessel and deployed as part of the Navy FMS solution.

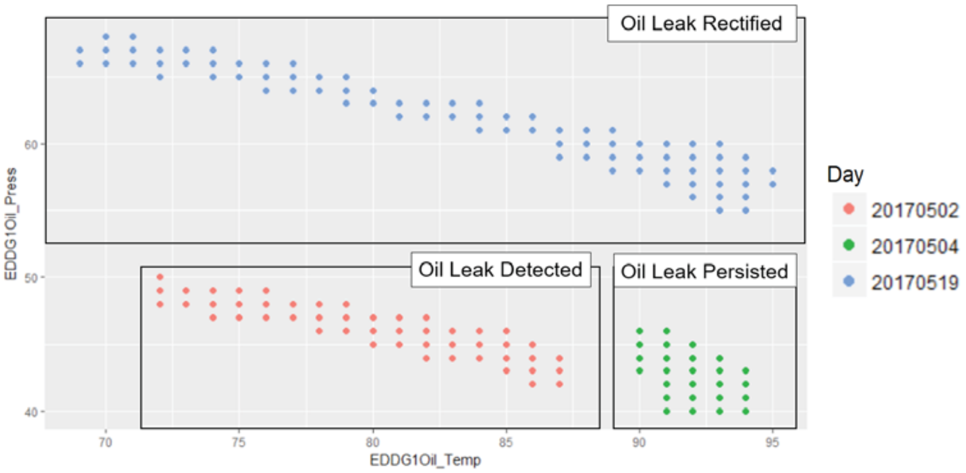


Figure 8. Scatter plot of the oil leak

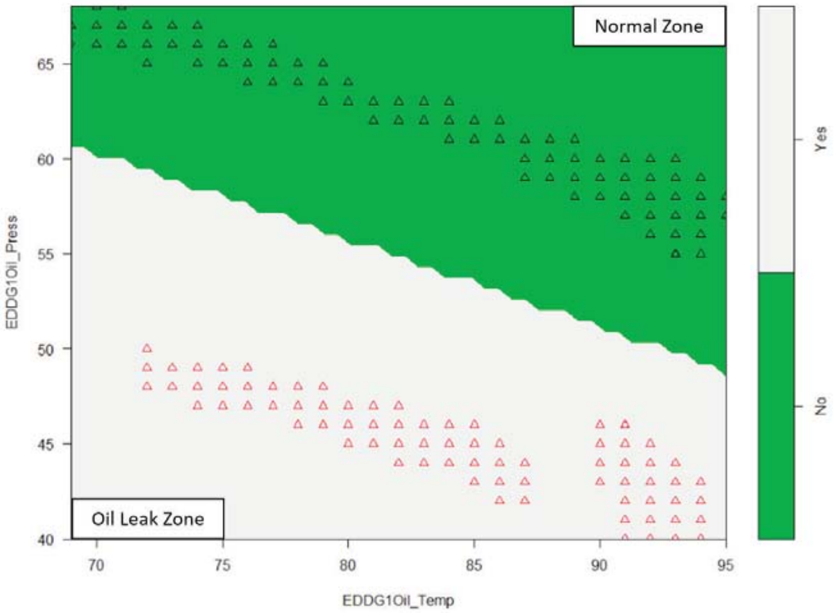


Figure 9. SVM classification plot

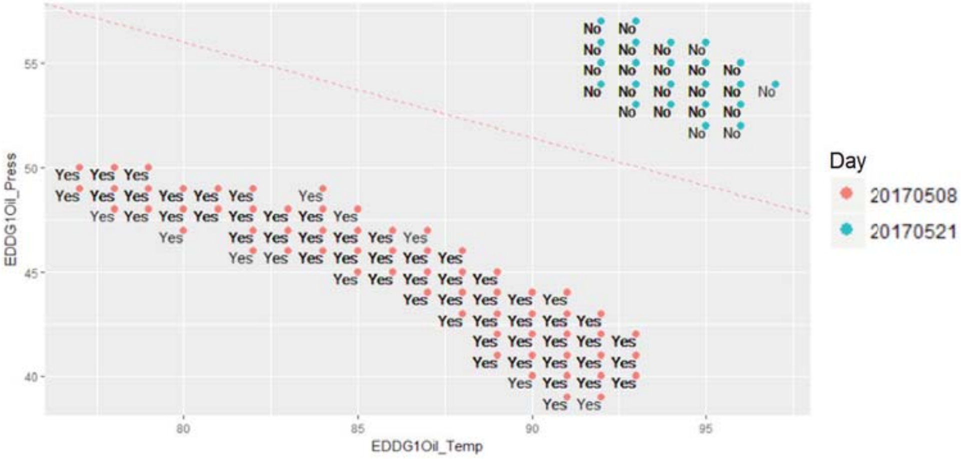


Figure 10. Validation plot of developed model

Data-Driven Governance in the NGPS

Analytics capabilities were developed for the NGPS to enhance procurement governance and management. Through trending and analysis of procurement data, DA enabled the detection of anomalies for proactive risk management and provided insights for better policy implementation. The following case study demonstrates the use of DA to detect related-party transaction.

Data Collection and Preparation

Related-party transaction is a fraud indicator that detects and highlights transactions that involve employees (such as requestors, approving officers, and tender evaluators) and vendors who are related, and cases where relatives of employees hold a stake in the vendor's company.

Relevant data were extracted from enterprise systems into the sandbox for analysis. Company registration and directorship data were also obtained from the Accounting and Corporate Regulatory Authority (ACRA).

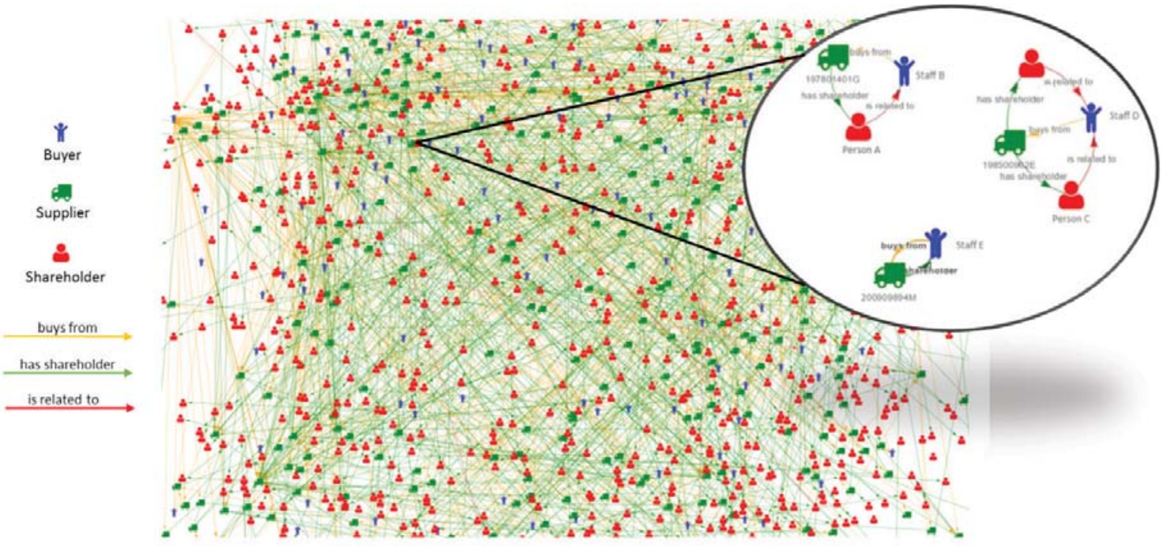


Figure 11. Relationship links of the relationship network

The following relationships were established from the data:

- Relationships between buyers, Approving Officers (AO), Quotation Requesting Officers (QRO) and suppliers were extracted from the Purchase Request and Order data.

- Relationships between buyers, AOs, QROs and their related parties were extracted from the enterprise data.

- Relationships between suppliers, their board members, owners and shareholders were extracted from the ACRA data.

These relationship data were used to generate a network analysis visualisation to facilitate the analysis. As shown in Figure 11, the network of relationships can be very complicated for human interpretation.

Applying Cycle Detection Technique

The Johnson's Algorithm was employed to detect suspicious relationships in the complex network. The analysis first broke the graphs down into strongly connected components as shown in Figure 12. Within each strongly connected component, the vertexes (the point where two or more curves, lines or edges meet) were ordered and an in-depth first search for every vertex was done to detect simple cycles.

The algorithm continued probing until a simple cycle was established when the start vertex and the end vertex became the same node. Figure 13 shows an illustration of a simple cycle detected from the strongly connected components. Using this model, the detected cyclic relationships in procurement transactions were flagged out for further investigation.

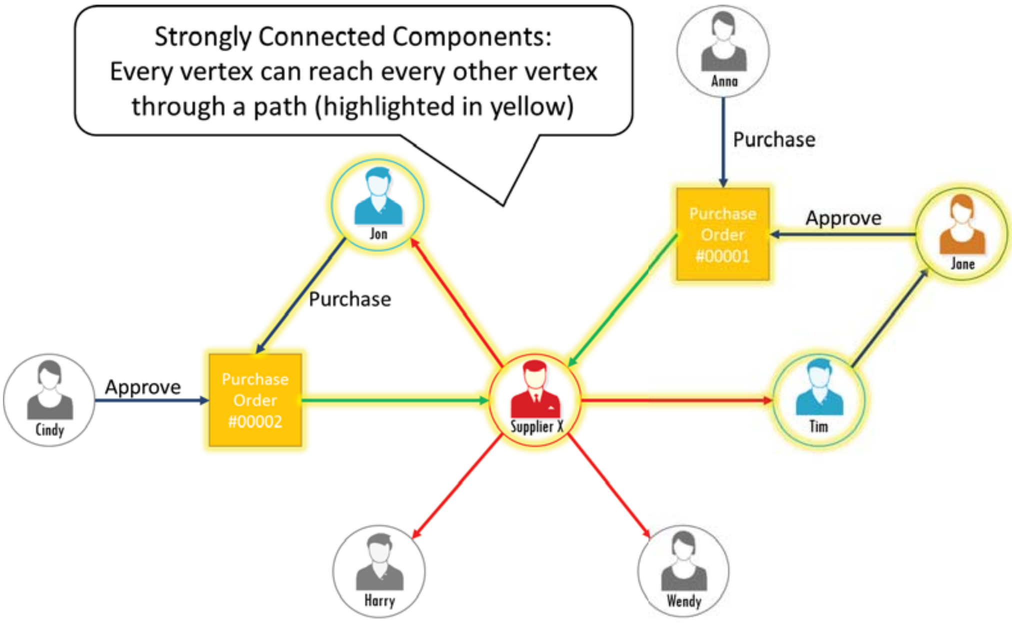

Strongly Connected Components: Every vertex can reach every other vertex through a path (highlighted in yellow)
Anna
Purchase
Purchase Order #00001
Approve
Jane
Jon
Purchase
Cindy
Approve
Purchase Order #00002
Supplier X
Tim
Harry
Wendy

Figure 12. Illustration of strongly connected components

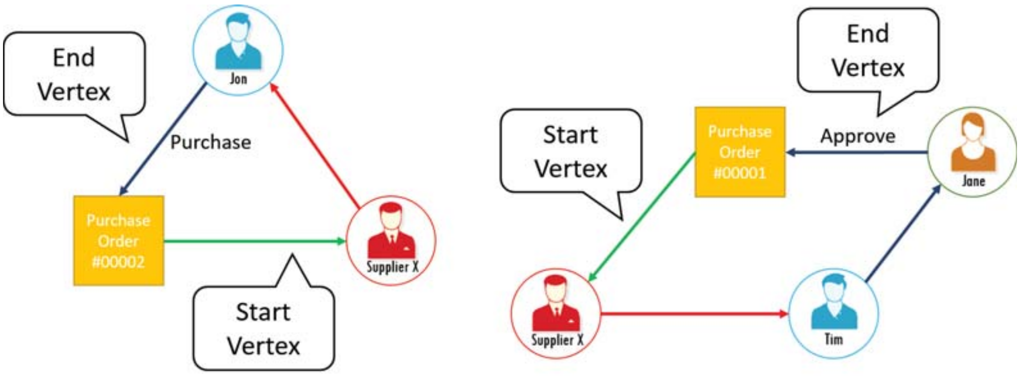


Figure 13. Illustration of simple cycles

CONCLUSION

An enterprise approach to put in place technology, data and people enablers provides a firm foundation for the development of DA capabilities in DSTA. This requires commitment from the senior management, good planning and innovative design to overcome challenges that are unique in DSTA's environment. The successful deployment of the EDAP and competency build-up of DSTA's engineers ensure a steady pipeline of new capabilities coming out from the lab to be deployed in production systems. The case studies for the FMS and the NGPS demonstrated the value of DA in the maintenance and procurement domains, among many other new capabilities such as information operations, social media analytics and video analytics that are being developed by DSTA.

ACKNOWLEDGEMENTS

The authors would like to thank Mr Koh Jin Hou and Mr Teo Hoon Beng for their help in organising, reviewing and editing the article.

REFERENCES

Cortes, C., & Vapnik, V. (1995, September). Support-vector networks. *Machine Learning*, 20(3), 273-297. doi: 10.1023/A:1022627411411

Dixon, J. (2010, October 14). *Pentaho, hadoop & data lakes*. Retrieved from https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

Johnson, D. B. (1975). Finding all the elementary circuits of a directed graph. *SIAM J. Comput, 4*(1), 77-84. doi: 10.1137/0204007

Russom, P. (2014). *Evolving data warehouse architectures in the age of big data*. Renton, Washington: TDWI Research.

Saaibi, S. P., & Diaz, J. P. M. (2013). *Understanding conflict of interest networks*. Retrieved from https://ethics.harvard.edu/blog/understanding-conflict-interest-networks

Srivastrava, A. N., & Han, J. (2012). *Machine learning and knowledge discovery for engineering systems health management*. Boca Raton, Florida: CRC Press.

Pogak, N., Prodanovic, M., & Green, T. C. (2007, March). Modelling, analysis and testing of autonomous operation of an inverter-based microgrid. *IEEE Transactions on Power Electronics*, 22(2), 613-625. doi: 10.1109/TPEL.2006.890003

BIOGRAPHY

HO Wee Peng Ricky is a Deputy Director at SAF C4 Command. Prior to his secondment, he was Programme Manager (Enterprise IT) working on the implementation of the Enterprise Data Analytics Platform (EDAP). He was also responsible for the development of the data analytics (DA) competency framework and training plan for DSTA. He has been contributing to the Singapore Government's Technical Reference Model, playing his roles as a member and deputy domain lead for Information Management since 2011. Ricky graduated with a Bachelor of Science degree with Honours in Computing Information Systems from the University of London in 1998.

KOH Lay Tin is a Senior System Architect (Enterprise IT) working on the implementation of the EDAP. She leads the implementation of DA projects in the Corporate IT domains such as procurement, audit, finance, medical and human resource. Lay Tin graduated with a Bachelor of Engineering (Electrical and Electronic Engineering) degree with Honours from Nanyang Technological University (NTU) in 1999. She also obtained a Master of Technology (Software Engineering) degree from the National University of Singapore (NUS) in 2004.

CHONG Yonghui Benjamin is a Principal Engineer (Enterprise IT) working on the implementation of the EDAP. Benjamin holds a Bachelor of Engineering (Electrical and Electronic Engineering) degree with Honours from NTU in 2006. He also obtained a Master of Technology (Enterprise Business Analytics) degree from NUS in 2016.

HO Chi Ming is a Principal Engineer (Enterprise IT) developing DA capabilities for the Fleet Management System. He enables technology push in DSTA by leveraging the Internet of Things and DA to assist the Singapore Armed Forces in streamlining their systems management. Chi Ming graduated with a Bachelor of Engineering degree with Honours in 2006 and a Master of Science in Business Analytics degree in 2016, both from NUS.