RAND CORPORATION

Research Report

WILLIAM MARCELLINO, JUSTIN GRANA, JAIR AGUIRRE, AMBER JAYCOCKS, CHRISTIAN JOHNSON, JOSHUA KERRIGAN

Army Analytic Capabilities

A Case Study Within Army Contracting Command and Its Implications

The U.S. Army faces two analytical and management challenges. The first is that its data are often locked away in siloed and proprietary databases. The second is that Army team members, such as data scientists, lack access to modern, common analytical tools. One potential solution to both problems is cloud migration: moving Army data to remotely accessed data environments offering scalable computer processing, data storage, and analytic services. To pilot the feasibility of this approach and gain insight, we developed a case study within Army Contracting Command (ACC) to see if there is a simple and effective way to overcome these challenges. We found that moving data to the cloud for analysis is effective and efficient and that there is a quick path forward to cost savings for the Army.

Although the specifics of this report are from within ACC, the implications are Army-wide. For example, in the remainder of this report, we discuss our effort within ACC and lessons learned, but this information also applies broadly to how the Army leverages insights from data. As an illustration of the power of robust data analytics, we also present findings from a demonstration machine-learning project to predict what kinds of contracts within ACC are likely to have unliquidated obligations (ULOs). The immediate findings are valuable and should be leveraged; our model performed well, reliably predicting contracts with ULOs with about 88-percent accuracy and identifying contracts with a potential reallocation value on the order of

KEY FINDINGS

- The U.S. Army can achieve immediate cost savings and efficiencies through advanced data analytics and the use of currently available commercial off-the-shelf technology.
- The Army does not need to wait for a complete system to reap efficiencies and cost savings. The Army can build off the proof of concept developed for this study.
- The Army can leverage commercial cloud infrastructure and software to immediately begin robust data sharing, querying, and analytics.
- Going to the cloud would provide infrastructure efficiently without large initial capital expenditures. Maintenance, upgrades, and hardware availability would be baked in.
- As a matter of policy, Army Contracting Command data scientists lack access to common data-science tools and lack permissions for remote access to computing infrastructure that allows for robust data-processing pipelines and analytic interfaces.

Abbreviations

ACC	Army Contracting Command
FPDS	Federal Procurement Data System
FY	fiscal year
GFEBS	General Fund Enterprise Business System
IT	information technology
OMB	Office of Management and Budget
PDF	Portable Document Format
POC	proof of concept
ULO	unliquidated obligation
VCE	Virtual Contracting Enterprise

$1 billion.[1] But we stress that the true significance of this pilot machine-learning effort is replication across the entire Army as an enterprise. Wherever the Army has data, and when those data become accessible, analytics can produce actionable insight for efficiencies and cost savings.

A Proof-of-Concept Study in Contract Analytics at Army Contracting Command

In fiscal year (FY) 2016, ACC awarded almost 170,000 contracts valued at $56.4 billion. ACC manages procurements through a process that includes requirements development, purchase request, solicitation, source selection, award, contract administration, and contract closeout. These procurements fall into several categories, such as systems, knowledge-based services, facilities and construction, equipment, ammunition and weapons, and research, development, test, and evaluation (RDT&E) appropriations.

ACC data scientists are limited by their information technology (IT) infrastructure and are unable to conduct analytics effectively on the basic set of structured contract data in the Virtual Contracting Enterprise (VCE).[2] Bandwidth and stability constraints mean they cannot query across all data adequately. ACC data scientists additionally lack IT permissions to develop analytic environments and install common analytic software tools on their local machines that would help extend the functionality of the current VCE infrastructure. Additionally, without a scalable data environment that enables remote access to personnel across the ACC, it is difficult for ACC data scientists to collaborate on data sets using a deep tool set. Instead, any activity is siloed computer by computer, and analytic resources are limited to what is available in the VCE.

Beyond infrastructure and access limits, there is a major technical hurdle in accessing and exploiting the majority of VCE data, which is in the form of unstructured, digitized documents—such as images in Portable Document Format (PDF)—that cannot be read by computers. Although existing structured metadata allow for some contract data to be accessed and managed in the primary database, many other contracts, particularly the larger ones that are vehicles for a diverse set of task orders, cannot be read or understood by computers. They are simply digitized images of contract documents and are not read as text by computers.

ACC needs the capacity to broadly manage its contracting enterprise, in particular to identify opportunities that would help the Army achieve cost savings and other efficiencies in its contracting enterprise. To help ACC leverage its existing domain expertise, RAND Arroyo Center conducted this study to demonstrate approaches for using analytics on ACC's vast stores of structured and unstructured contracting data.

Access to Unstructured Contract Data

Perhaps the biggest barrier that ACC faces right now in contract analytics is that the majority of

[1] Here, we define *accuracy* as the number of correct (true positive and true negative) predictions divided by the total number of predictions. The model additionally had a precision rate of 68 percent and recall of 82 percent.

[2] The VCE is a suite of web-based contracting tools that provides standardization across the enterprise while providing oversight for all contracting and workforce data. For more information, see U.S. Army, "Virtual Contracting Enterprise," webpage, undated.

contract data, including granular information that would most inform deep analytics, is *unstructured*. This barrier is likely a problem faced more broadly throughout the Army. Structured data are set up to be machine readable like databases and tabular information. Unstructured data, however, are not set up for machines to read—for example, a word that is spelled out fully and abbreviated or misspelled. An additional barrier is that this unstructured data (mostly PDFs but also emails and spreadsheets) have not been converted into text: They are still essentially images of text.

As a proof of concept (POC) for overcoming these barriers, we piloted an effort to make the unstructured contract data machine readable for querying, easy retrieval, and analytics. Nine years of contract data—about 500,000 contracts—were extracted from the proprietary archive format they are stored in. We then used Apache Solr, an open-source technology, to index those files. Conceptually, this is akin to filing and labeling a giant mass of documents for quick retrieval later. With documents now available on demand, we then used open-source optical character recognition software to extract the text from these documents in a computer-readable format, allowing for text analytics on contract data.

Using this open-source technology as a base, we were able to search for and access contracts using a simple graphical interface and plain-language search terms, similar to Google. An analyst could search for "'cut' AND 'grass'" and get back results, and then see that "groundskeeping" is an additional useful search term.

A Way Forward to the Cloud

Cloud computing was essential to our project for several reasons. The VCE currently lacks a scalable analytic infrastructure, and building one would involve a prohibitive capital investment. Leveraging an existing cloud-computing service meant scalability, flexibility, agility, and data security. Furthermore, the cost for these tools is based on storage and usage, making cloud-based services a cost-effective solution because there are no large up-front investments in software licenses.

This move to the cloud yielded four sets of insights that we think are valuable to the Army as it moves toward adopting data analytics tools more broadly. These issues apply to IT permissions, the use of cloud infrastructure to increase collaboration and reduce duplication of effort, technical insights to inform next steps, and the principles that guide matching solutions to needs. We were able to overcome multiple hurdles to move ACC data to a robust cloud environment, and the lessons gleaned from this effort will be valuable across the Army's enterprise effort to build contract analytics capacity.

ACC Data Scientists Do Not Have Adequate Access to Analytic Tools

In the course of our research, we encountered an unexpected tension between data security and data analytics. Within ACC, restrictions exist on the installation and use of software and software libraries that have not been pre-approved for use on government-managed hardware and in particular, tools that allow users to access remote computing environments and interact with data stored remotely. These restrictions are not arbitrary; there is a legitimate need to secure Army data and networks. However, data scientists present a unique use case in that they have a legitimate need to access software and remote databases in order to help the Army save money by exploiting its data. ACC can work with the G-6 to extend network access and software installation permissions to those outside of IT roles, especially data scientists and analysts. Additionally, the G-6 could be involved in reviewing and curating permissions for necessary software and tools.

A Cloud Infrastructure Enables Collaboration and Reduces Duplication of Effort

Regardless of particular solutions, exposing data for collaboration is important. ACC can benefit by maintaining a single foundational data source of contracts and contract data that is available to all relevant personnel. Additionally, ACC can benefit by maintaining a dedicated cadre and process for

managing and cleaning data. This can enable a collaborative analytic environment where multiple organizations can contribute by building on previous efforts. Practically speaking, this means providing access to a common flexible and scalable computing and storage environment, such as a public, private, or hybrid cloud.

Lessons Learned Can Help with Technical Challenges in Cloud Analytics

Because contracting data sets can be large, cloud analytics require robust connectivity and bandwidth that may need to be provisioned. Establishing best practices in transferring data, for example, can help reduce networking and processing loads. Additionally, skilled administrators who can implement coordinated sets of services and control permissions based on users' business functions and skill levels are also required. Competency in such functions across an organization requires standardization of routine tasks as well as documentation of heroic first-of-their-kind efforts.

Guiding Principles for Matching Cloud-Based Solutions to Needs

Given the number of options for data environments and analytic capabilities, we suggest carefully documenting needs prior to selecting and configuring specific solutions and platforms. We found that, because different users have different needs, ACC should document needs and select matching options—for example, advanced database capabilities, visualization and querying capabilities, and machine-learning capabilities.

Infrastructure as code—characterized by turning processes and infrastructure into code that can be copied and pasted into new settings—will also be important. Converting successful work into code stores the effort and makes it instantly portable for follow-on application, with no need for duplicating the prior effort. Both analytic tools and environment infrastructure could also be stored, managed, and shared as code.

Finally, the trade-off space between open-source and proprietary software is complex. In conducting contract-analytics market research, it was clear that vendors claim that their proprietary solutions offer superior performance and thus are worth added expense. Furthermore, proprietary solutions can be tailored to enterprise specifics and may come bundled with significant support options. However, we experienced firsthand how proprietary software could inhibit collaboration. For instance, we had to invest considerable effort and expertise to convert data from a proprietary format and make it accessible. Data access is critical to contract analytics, but if any party does not have access to the proprietary software, they are excluded from the effort. This is a particular issue for collaboration with external parties. In contrast, open-source technologies and software offer considerable power and can enhance collaboration, as this POC demonstrates. In addition to cost savings, choosing open-source solutions can work to serve the principle of data access. The Army should conduct a cost-benefit analysis of the trade-offs between proprietary and open-source software solutions.

Phased Approach to Increasing Data Analytics Capability for ACC

This POC effort establishes a foundation for the three critical dimensions of cloud analytics: moving data, the data environment, and analytics capability connected to the environment. Given this foundation and the recommendations from the study, ACC—and eventually the Army as an enterprise—is positioned to build on this foundation using a phased implementation approach (see Figure 1).

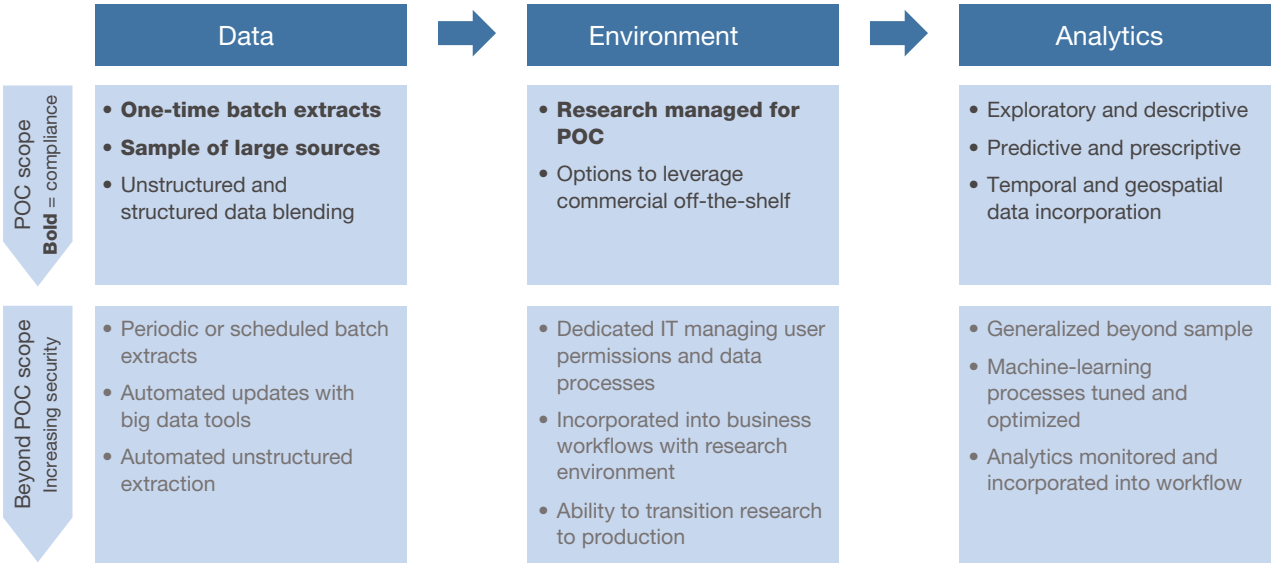
By following these recommendations, the Army can quickly build up a robust contract-analytics capability that potentially generates cost savings and operational efficiencies.

How Data Analytics Can Unlock Cost Savings and Efficiencies

The most important finding in this study was the potential scope and impact of contract analytics in

FIGURE 1

A Phased Technology Approach



the near term for the Army. Even simple relatively descriptive analytics and machine-learning applications could add value to the enterprise. As an illustrative example, we team built a machine-learning model to identify contracts with ULOs at the end of the fund's life cycle, when the fund is canceled and is no longer available for the Army to use. This model performed well, reliably predicting contracts with ULOs with about 88 percent accuracy and with a potential reallocation value on the order of $1 billion. Perhaps more important, by identifying the crucial predictors of ULOs, this effort shows how ACC can change policy to proactively be more efficient in contract funding.

Efficient contract management includes effective allocation of contracting funds. However, proper allocation of funds is not always possible, as incorrect estimates of either contract duration or cost can ultimately leave a contract with unexhausted funds. For example, contracts with ULOs reduce the number of additional contracts that can be funded. Thus, conducting a ULO-focused predictive analysis on all of the Army's contracts can help serve two valuable and complementary purposes: identifying contracts that are likely to result in ULOs and determining the factors that are strong predictors of contract ULOs. *An important finding in this study was that even relatively simple descriptive analytics and machine-learning applications developed and deployed in the right analytic environment can help the Army quickly identify cost-savings opportunities*, such as why some contracts result in ULOs.

A Machine-Learning Model to Predict Contracts with ULOs

Using General Fund Enterprise Business System (GFEBS) and Federal Procurement Data System (FPDS) contract data over FYs 2012–2014, we pulled 300,000 contracts with 150 features describing various aspects of each awarded contract, including financial data, chronological events, and industry information. ULOs in this data set vary between $0 and $100,000,000, where contracts with zero-valued ULO are the norm and nonzero ULO represent about 25 percent of the contracts.

We used a well-established, relatively simple shallow algorithm (random forest) classifier on the data to build a predictive model for contacts with ULOs.[3] This model performed well, reliably predicting contracts with ULOs with about 88

[3] Andy Liaw and Matthew Wiener, "Classification and Regression by randomForest," *R News*, Vol. 2/3, December 2002.

percent accuracy and with a potential reallocation value on the order of $1 billion. Most important, by identifying critical predictors of ULOs, ACC can be proactive and more efficient in contract funding. Table 1 presents the most important contract features in predicting ULOs, in descending order of importance.

To help explain the value of proactively identifying ULO indicators, the following figures and tables show three illustrative, easy-to-understand indicators of higher ULO contracts. For example, *contract duration* is a strong predictor of ULOs. Using the results of the model, we mapped the duration of a contract in years to predict the probability of ULO. As contracts get longer, the likelihood of ULO tends to increase until it tapers off after two years and then begins to increase again at around three and a half years. Figure 2 shows this effect and indicates where additional attention should be given to contracts.

Given that the algorithm showed that the funds center is also a strong predictor of ULOs, we further sought to understand which Funds Centers have the highest prevalence of ULOs. As Table 2 shows, Research, Development and Engineering Command, for example, has the highest number of ULOs contracts, whereas Aviation and Missile Command has the highest likelihood (about 60 percent) of ULOs.

In a final example, we show how OMB descriptions predict likely contract ULOs (Table 3). Although "equipment" is the highest absolute number for ULO contracts, "advisory and assistance services" had the highest likelihood of ULOs with approximately 43 percent.

If Army Materiel Command wants to prevent ULOs, understanding the contract features that have an increased association with ULOs should be the first step.

TABLE 1

GFEBS-FPDS Contract Feature Names and Descriptions in Descending Order of Importance

Feature Name	Description
Fund	Application of funds alphanumeric code
Funds Center	Alphanumeric value for organizational element that receives, distributes, and manages funds
Contract duration	Duration of contract from effective date to completion date
Purchasing document type	Document format for the contract
Contracting office name	Command which sponsored the contract
Purchasing group	Contract customer or buyer
North American Industry Classification System description	Industry-specific description of service, part, or equipment
Portfolio	Industry group for contract
Current contract value	Contract value as of the current date
Office of Management and Budget (OMB) object class description	Description of the service, part, or equipment
Ultimate contract value	Contract value over duration of contract
Contract instrument type	Type of Federal Acquisition Regulation contract

NOTE: These variables were selected out of about 150 total variables because they each provided an appropriate (more than two and fewer than about 20) different categorical entry to describe each contract.

FIGURE 2

Probability of ULO by Contract Duration

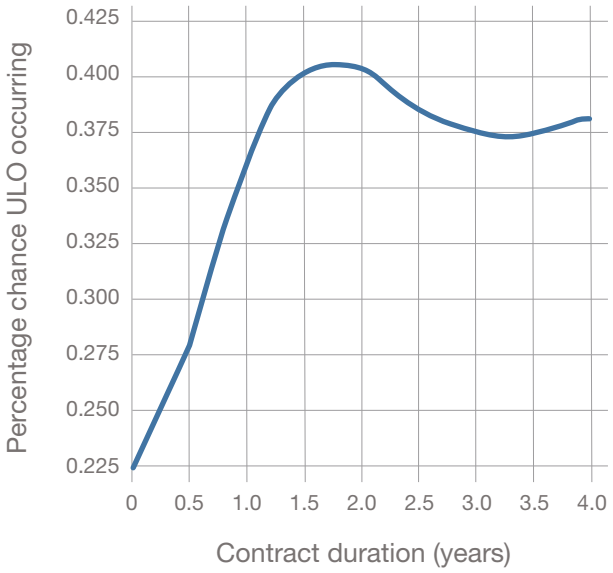


TABLE 2

ULO Percentages by Funds Center

Funds Center	Description	*N*	ULO Percentage
A60F*	Research, Development and Engineering Command	16,816	29.1
A60X*	Aviation and Missile Command	8,892	60.0
A5XF*	Space and Missile Defense Command	8,499	55.1
A60D*	Communications-Electronics Command	8,080	55.8
A5XE*	Program Executive Office, Command Control Communications–Tactical	4,531	42.5
A5XH*	Program Executive Office, Intelligence, Electronic Warfare and Sensors	3,962	37.7
A2AI*	Installation Management Command, Headquarters Army Reserve Division	2,559	54.4
A5XD*	Program Executive Office, Combat Support and Combat Service Support	2,336	11.9

NOTE: Funds Center codes are truncated (indicated with the asterisk) at the coarsest organization level to provide as much unique information as possible without being exhaustive.

TABLE 3

ULO Percentages by Object Class

OMB Object Class Description	*N*	ULO Percentage
Equipment	30,246	9.7
Advisory and assistance services	29,587	42.8
Supplies and materials	20,471	11.1
Operation and maintenance of facilities	19,206	9.0
Operation and maintenance of equipment	13,931	37.7
Research and development	12,053	26.9
Communications, utilities, and miscellaneous charges	4,981	22.7
Land and structures	2,296	2.1
Other services from nonfederal sources	1,771	10.6
Rental payments to others	1,531	6.5
Travel and transportation of persons	1,122	7.2

Conclusion: The Army Can Immediately Benefit from Cloud Analytics

This report outlined a set of challenges around data access, specifically within ACC but relevant to the entire Army. We laid out the results of a pilot study within ACC, which migrated data to the cloud, and demonstrated how a cloud data environment can solve a number of access challenges. Furthermore, we showed how a relatively simple machine-learning pilot effort identified an immediate $1 billion opportunity to improve contract funding misallocations. Most important, we believe immediate benefits are available; the Army does not need to wait for a complete solution to gain efficiencies and save money. An agile, immediate effort will have enormous value across the Army anywhere that data are available.

Recommendations

Based on our pilot effort, we make the following recommendations.

1. To validate the methodology from this report across multiple commands, the Army should immediately conduct multiple similar POCs that take siloed and inaccessible data to the cloud to be analyzed using modern analytical tools.
2. The Army should develop a policy on the use of open-source analytical products and cloud-storage requirements to ensure that multiple ongoing data efforts are interoperable.
3. The Army should set a goal, perhaps not more than one year out, to have access to a scalable analytics environment, such as the one described in this report, for all of its key operational and business data.

References

Liaw, Andy, and Matthew Wiener, "Classification and Regression by randomForest," *R News*, Vol. 2/3, December 2002. As of November 22, 2019: https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf

U.S. Army, "Virtual Contracting Enterprise," webpage, undated. As of June 10, 2020: https://vce.army.mil/Portal/

About This Report

The research reported here was completed in July 2020, followed by security review by the sponsor and the Office of the Chief of Public Affairs, with final sign-off in March 2021.

This report documents research and analysis conducted as part of the project *Army Contract Analytics Capability Development*, sponsored by U.S. Army Materiel Command. The project aimed to help the U.S. Army develop the capability to analyze all data available in the Virtual Contracting Enterprise, with a focus on extracting information from unstructured text documents and combining it with other structured data to find opportunities for cost savings. A longer and more technically oriented version of this report—containing information on research tools, analytic workflow diagrams, performance metrics, and previous work—has been documented by the authors of this report in unpublished RAND Corporation research in 2020.

This research was conducted within RAND Arroyo Center's Forces and Logistics Program. RAND Arroyo Center, part of the RAND Corporation, is a federally funded research and development center (FFRDC) sponsored by the United States Army.

RAND operates under a "Federal-Wide Assurance" (FWA00003425) and complies with the Code of Federal Regulations for the Protection of Human Subjects Under United States Law (45 CFR 46), also known as "the Common Rule," as well as with the implementation guidance set forth in DoD Instruction 3216.02. As applicable, this compliance includes reviews and approvals by RAND's Institutional Review Board (the Human Subjects Protection Committee) and by the U.S. Army. The views of sources utilized in this study are solely their own and do not represent the official policy or position of DoD or the U.S. Government.

Acknowledgments

The authors thank U.S. Army Materiel Command for sponsoring this work and Christopher Hill, director, Army Materiel Command Analysis Group at U.S. Army Materiel Command, for his interest and support of the project. Thanks also go to Kevin Foster, chief of the Data Analytics Division at Army Contracting Command, for his help coordinating and supporting the various stakeholders and data sources needed for this project. Thanks also go to Peter Schirmer of RAND and MG John Ferrari, U.S. Army (Ret.), for their invaluable advice and insight on this report. Finally, the authors thank Shawn McKay of the RAND Corporation for his help and support piloting machine learning approaches to contracting data.

For more information on this publication, visit www.rand.org/t/RRA106-1.

www.rand.org

RR-A106-1