SR-21-003

Range Commanders Council

DATA SYSTEMS GROUP

TRMC BIG DATA ANALYTICS ARCHITECTURE ASSESSMENT

ABERDEEN TEST CENTER
DUGWAY PROVING GROUND
ELECTRONIC PROVING GROUND
REAGAN TEST SITE
REDSTONE TEST CENTER
WHITE SANDS TEST CENTER
YUMA PROVING GROUND

NAVAL AIR WARFARE CENTER AIRCRAFT DIVISION PATUXENT RIVER
NAVAL AIR WARFARE CENTER WEAPONS DIVISION CHINA LAKE
NAVAL AIR WARFARE CENTER WEAPONS DIVISION POINT MUGU
NAVAL SURFACE WARFARE CENTER DAHLGREN DIVISION
NAVAL UNDERSEA WARFARE CENTER DIVISION KEYPORT
NAVAL UNDERSEA WARFARE CENTER DIVISION NEWPORT
PACIFIC MISSILE RANGE FACILITY

96th TEST WING
412th TEST WING
ARNOLD ENGINEERING DEVELOPMENT COMPLEX

SPACE LAUNCH DELTA 30
SPACE LAUNCH DELTA 45

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

DISTRIBUTION A: APPROVED FOR PUBLIC RELEASE
DISTRIBUTION IS UNLIMITED

This page intentionally left blank.

SR-21-003

TRMC BIG DATA ANALYTICS ARCHITECTURE ASSESSMENT

September 2021

Prepared by

Data Sciences Group

Published by

Secretariat
Range Commanders Council
US Army White Sands Missile Range
New Mexico 88002-5110

This page intentionally left blank.

Table of Contents

This page intentionally left blank.

Preface

This document is the result of a task for the Data Sciences Group to formally assess the Test Resource Management Center Big Data Analytics Architecture as documented in the "Knowledge Management and Big Data Analytics Architecture Framework" document and to provide a recommendation as to whether the architecture should be adopted as an RCC standard.

An enterprise approach for knowledge management and big data analytics would provide for more robust and reliable data storage and greater discoverability, accessibility to this test data, and significantly improved ability to perform large-scale data analysis. These would result in improved test and evaluation of Department of Defense weapon systems, leading to better weapon systems for the warfighter. An effective enterprise approach requires a well-designed architecture, and all participating ranges would benefit by providing input on the architecture.

This assessment was conducted by the Range Commanders Council Data Sciences Group (RCC DSG) in support of RCC Task DS-014, and will be presented to TRMC for their consideration upon publication by the RCC Secretariat.

For questions regarding this document, contact the RCC Secretariat at:

Secretariat, Range Commanders Council
ATTN: TEWS-TDR
Building 1510 Headquarters Avenue
White Sands Missile Range, New Mexico 88002-5110
Telephone: (575) 678-1107, DSN 258-1107
E-mail: rcc-feedback@trmc.osd.mil

This page intentionally left blank.

Acronyms

BDAA	Big Data Analytics Architecture
DSG	Data Sciences Group
RCC	Range Commanders Council
RDBMS	relational database management systems
TRMC	Test Resource Management Center

1. Introduction

This document is an assessment of the Test Resource Management Center (TRMC) Knowledge Management and Big Data Analytics Architecture (BDAA) framework as documented by the TRMC.[1]

For simplicity of reference, the above TRMC (2019) publication will be referred to in this assessment as the TRMC BDAA.

As an important point of clarification, this assessment focused on the general architectural framework documented in the TRMC BDAA and did not focus on any specific implementation of the TRMC BDAA.

The organization of this assessment is as follows.

1) In Section 2, brief overview of the TRMC BDAA will provide the context for this assessment. In addition to the overview, it is assumed the reader of this assessment will have a copy of the TRMC BDAA.
2) Section 3 provides the reviewer comments on the TRMC BDAA. These comments will be grouped into sections according to the TRMC BDAA chapter they are applicable to.
3) Section 4 provides specific recommendations for the TRMC BDAA. These recommendations will be based upon a consolidation of reviewer comments that address common areas of concern.

2. Architecture Overview

The following is an extended excerpt of the executive summary from the TRMC BDAA, which serves as a good overview of the analytics architecture.

> The Big Data Analytics Architecture (BDAA). The purpose of this document is to put forward a broad architectural framework (the BDAA) for use as a template in the T&E community for creating a next generation analytics capability, called the Big Data Analytics System (BDAS). The architecture specifies general T&E requirements for big data analytics, as well as certain goals and constraints for security, software, and hardware in the future BDAS. From a software perspective, the architecture specifies a reliance on mostly open-source software that has been developed by both government and industry. From a hardware perspective, the architecture envisions additional storage and processing deployed to the test ranges, connected up to more centralized "cloud-based" Regional Analytic Capabilities. From a security perspective, multiple options are presented that allow each range to define the security boundaries where they feel most comfortable according to their interpretation of the Risk Management Framework, either mostly centrally controlled, mostly locally controlled, or a hybrid approach. The architecture envisions a governance process based on a partnership between the individual ranges,

[1] Test Resource Management Center. *Knowledge Management and Big Data Analytics Architecture Framework*. Version 13. 31 January 2019. May be superseded by update. Retrieved 23 July 2021. Available to people with TRMC credentials at https://www.trmc.osd.mil/wiki/display/JMETC/Big+Data+Knowledge+Management?preview=%2F55968739%2F55968745%2FBigDataArchitecture-v13-2019-01-31-DistA.pdf.

Services, and the TRMC, with standardization of architecture elements undertaken by a Configuration Control Board consisting of user organizations.

A conceptual overview of the BDAA is shown in Figure 1, below.

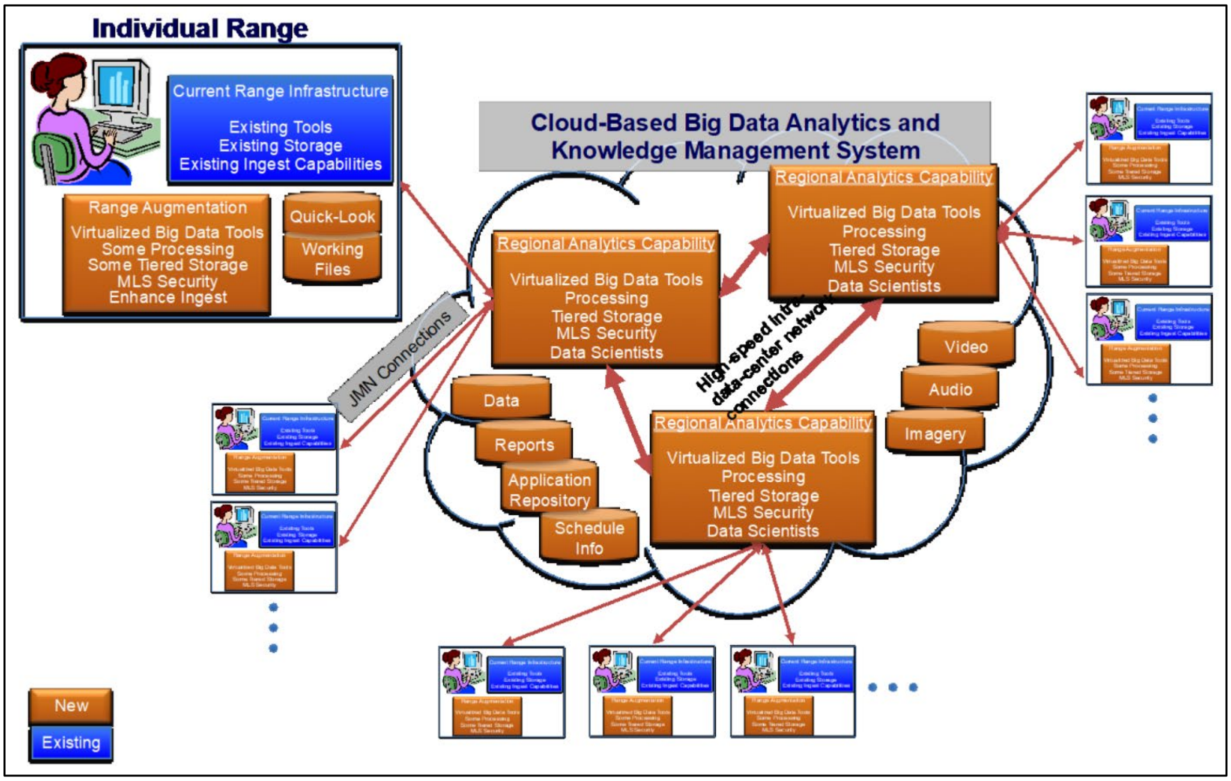

Figure 1. Big Data Analytics Architecture Overview

Each range has its existing infrastructure and tools, augmented by some additional storage, processing, visualization, security, and network capabilities; and is connected to the set of Regional Analytics Capabilities (RACs) via the Joint Mission Environment Test Capability (JMETC) Multiple Independent Levels of Security (MILS) Network (JMN). Initial data processing and analysis are done at individual ranges; then, when appropriate, data is migrated to the RACs where additional analysis across many programs and test events can be performed. The ranges and the RACs together represent one integrated federated data management system, so that any analyst with the correct access permissions can perform analysis across all of the data in the entire enterprise, no matter where any individual piece of data actually resides. The architecture provides for data redundancy, continuity-of-operations, and high reliability of data storage and access. (TRMC BDAA, pages 2-3)

3. Reviewer Comments

The TRMC BDAA is divided into the following major sections.

1) Executive Summary
2) Vision and Requirements
3) Knowledge Management and Big Data Analytics

4) Software and Data Architecture
5) Hardware Architecture
6) Security Architecture
7) Architecture Development and Maintenance
8) Appendices

The reviewer comments are grouped into similar sections according to the TRMC BDAA section they are applicable to. Reviewer comments identifying spelling, grammatical, or typographical errors were not included in this assessment document.

Each reviewer comment is tagged with one or more keywords to identify the central issue or issues described by the comment. These comment tags are then used to consolidate the comments into recommendations for improving the TRMC BDAA. Each comment tag per reviewer comment is identified on a separate line and prefixed with "CONCERN:" in order to streamline searching within this document.

Note that when reviewers are referencing sections of "this document", these are references to the TRMC BDAA document and not to this assessment document.

3.1 Executive Summary

3.1.1 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: NONE

No specific comments, other than ensure that the Executive Summary is updated as necessary in order to reflect any changes made to the core document itself based upon the comments received related to the core document itself.

3.2 Vision and Requirements

3.2.1 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: DATA VISIBILITY

The document discusses both broad sharing of data and the ability for ranges/programs to keep their data private. With such a system, presumably there would need to be a mechanism/procedure for others to request access to private data, but I couldn't find anything in the document about "advertising" or anything like that. I would expect that to be under the requirements section, if only to drive home that, yes, you can keep your data private, but you have to at least let others know what you have, in case it could be useful to them.

3.2.2 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]

COMMENT TAG: CONCERN: COST
In Section 1.1 the following statement is made:

> The amount of data collected in testing has grown beyond the capacity of current analytic tools and methods to make proper use of the data. It is almost impossible to examine previously collected data and compare it with recently collected data due to the lack of adequate on-line storage. New technological approaches, collectively known as BDA, offer an opportunity to mine collected data to gain new insights, reduce the number of tests and therefore the overall testing cost, and reuse testing data. [Section 1.1, page 6]

It is entirely possible that while BDA may reduce the number of tests, BDA may actually increase the overall testing cost. This is because the additional infrastructure costs incurred for the additional storage and computing and network resources needed to enable BDA, along with the additional personnel costs for the additional system administrator and data scientist technical expertise needed to enable BDA, may be greater than the cost reductions realized due to the reduced number of tests. Hence, cost savings by using BDA are certainly not guaranteed and perhaps not even likely, and consequently cost savings should not be identified as a primary driver for utilizing BDA.

The primary drivers for utilizing BDA should be: (a) the ability to answer questions faster (i.e., faster analysis), and consequently (b) the ability to answer questions that are not practical to answer without BDA because it would take too long to answer them (i.e., better analysis). If we can save money with BDA then great, but we should be willing to accept and perhaps expect that it will cost us more money to squeeze more information faster out of the data we collect.

Recommend rewriting the last sentence of the STATEMENT above to be: "New technological approaches, collectively known as BDA, offer an opportunity to mine collected data to gain new insights, reduce decision making time, and improve evaluation quality."

3.2.3 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: NONE

Given the central importance of both "knowledge management" and "big data analytics" to this document, recommend changing the title of this section to "The Value of Knowledge Management and Big Data Analytics".

Editor's Note: The "this section" referred to above is Section 1.2 of the TRMC BDAA.

3.2.4 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: NONE
In Section 1.2 the following statement is made:

> In this document, the term "knowledge management" refers to the software and storage systems designed to store, reconstruct, retrieve, display, and interrelate all of the data, conclusions, and analysis produced during the T&E process. The term "data" relates to all three terms: data, information, and knowledge. Knowledge management is discussed in more detail in Section 2.1, below. Related to knowledge management is the term "big data

> analytics," which deals with how analysis is done on datasets that are too large to analyze on a single computer. Big Data Analytics is described in more detail in Section 2.2, below. For T&E data, both knowledge management and big data analytics are interrelated and required together - neither can be fully realized without the other. [Section 1.2, page 7]

Given the central importance of "knowledge management" (KM) and "big data analytics" (BDA) to this document, recommend putting this statement into its own paragraph rather than keeping at the end of a larger paragraph.

3.2.5 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: COST
In Section 1.2 the following statement is made:

> Transforming the current T&E data management infrastructure to one employing a KM/BDA approach will support both current warfighter T&E needs and the developmental and operational testing of future weapon platforms. The knowledge management enterprise will greatly enhance the T&E value proposition for all respective users and stakeholders. The T&E community will be able to realize improvements in cost avoidance and cost reductions, in faster and more accurate T&E responses, and in overall T&E capabilities. [Section 1.2, page 7]

It is entirely possible that while BDA may reduce the number of tests, BDA may actually increase the overall testing cost. This is because the additional infrastructure costs incurred for the additional storage and computing and network resources needed to enable BDA, along with the additional personnel costs for the additional system administrator and data scientist technical expertise needed to enable BDA, may be greater than the cost reductions realized due to the reduced number of tests. Hence, cost savings by using BDA are certainly not guaranteed and perhaps not even likely, and consequently cost savings should not be identified as a primary driver for utilizing BDA.

The primary drivers for utilizing BDA should be: (a) the ability to answer questions faster (i.e., faster analysis), and consequently (b) the ability to answer questions that are not practical to answer without BDA because it would take too long to answer them (i.e., better analysis). If we can save money with BDA then great, but we should be willing to accept and perhaps expect that it will cost us more money to squeeze more information faster out of the data we collect.

Recommend changing the bullets in this section from "Cost Reduction, Cost Avoidance, Timely Response, Capabilities" to "Faster Analysis, Better Analysis".

3.2.6 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: COST
COMMENT TAG: CONCERN: OTHER ARCHITECTURES/GUIDANCE
In Section 1.3 the following statement is made:

> TRMC's KM/BDA vision is to build a DoD T&E knowledge management and analysis capability that leverages commercial big data analytic and cloud computing technologies

> to improve evaluation quality, reduce decision making time, and reduce T&E cost. This vision encompasses a big data architecture framework—its supporting resources, methodologies, and guidance—to properly address the current and future data needs of weapon systems testing. The architecture itself is called the Big Data Analytics Architecture (BDAA). The system built in conformance to the architecture is called the Big Data Analytics System (BDAS). The commercial technologies included comprise commodity hardware and open source software, which makes the required scaling both affordable and achievable. [Section 1.3, page 8]

While BDA can certainly provide faster analysis and better analysis, there is no guarantee that BDA can do this at cheaper cost. This is because the additional infrastructure costs incurred for the additional storage and computing and network resources needed to enable BDA, along with the additional personnel costs for the additional system administrator and data scientist technical expertise needed to enable BDA, may be greater than the cost reductions realized due to the reduced number of tests. Hence, cost savings by using BDA are certainly not guaranteed and perhaps not even likely, and consequently cost savings should not be identified as a primary driver for utilizing BDA.

The primary drivers for utilizing BDA should be: (a) the ability to answer questions faster (i.e., faster analysis), and consequently (b) the ability to answer questions that are not practical to answer without BDA because it would take too long to answer them (i.e., better analysis). If we can save money with BDA then great, but we should be willing to accept and perhaps expect that it will cost us more money to squeeze more information faster out of the data we collect.

Recommend rewriting the first sentence of the STATEMENT above to be:

> "TRMC's KM/BDA vision is to build a DoD T&E knowledge management and analysis capability that leverages commercial big data analytic and cloud computing technologies to reduce decision making time and improve evaluation quality."

Given that KM and BDA are separate but interrelated concepts, is there also a "Knowledge Management Architecture (KMA)" that the BDAA will need to interface with, and if so then how? Recommend additional discussion and clarification on the relationship between KM and BDA, and how the BDAA interacts with the KMA (if it exists) to meet the TRMC KM/BDA vision.

Also, the National Institute of Standards and Technology (NIST) published a Big Data Reference Architecture in September 2015.[2] Recommend additional discussion on how the BDAA relates to the NIST Big Data Reference Architecture.

3.2.7 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: KM/BDA/REQUIREMENTS

According to the Executive Summary, the purpose of this document is to describe the BDA Architecture (BDAA). According to Section 1.2 of this document, KM and BDA are separate but interrelated concepts. Table 1 in Section 1.4 of this document is titled "T&E KM Enterprise High-Level Requirements." Are the BDAA requirements synonymous with the T&E KM Enterprise

[2] National Institute of Standards and Technology. *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*. SP 1500-6. September 2015. Retrieved 23 July 2021. Available at http://dx.doi.org/10.6028/NIST.SP.1500-6.

requirements? If yes, then why not just title Table 1 as “BDAA High-Level Requirements”? If no, then where are the BDAA high-level requirements? Should the BDAA be renamed the KM/BDA Architecture? Recommend additional discussion and clarification on the relationship between KM and BDA.

3.2.8 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: COMMODITY HARDWARE/OPEN SOURCE SOFTWARE

Recommend adding the following bullet to Table 1 under the Fundamental Function Ingest: “Commodity hardware (i.e., easily available, relatively inexpensive, and interchangeable with other hardware performing the same function) and open-source (i.e., non-proprietary) software meeting data ingest requirements shall be used whenever possible.”

3.2.9 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: COMMODITY HARDWARE/OPEN SOURCE SOFTWARE

Recommend adding the following bullet to Table 1 under the Fundamental Function Warehouse: “Commodity hardware (i.e., easily available, relatively inexpensive, and interchangeable with other hardware performing the same function) and open-source (i.e., non-proprietary) software meeting data storage requirements shall be used whenever possible.”

3.2.10 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: DATA SOURCES

In Section 1.4 the following statement is made:

> The enterprise shall be capable of ingesting from both traditional and non-traditional data sources. [Section 1.4, Table 1, Fundamental Function Ingest, page 10]

It is not clear what is considered a “traditional” data source or a “non-traditional” data source. Recommend providing a few examples of each to clarify the distinction.

3.2.11 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: INGEST

In Section 1.4 the following statement is made:

> The enterprise shall be able to ingest all relevant data types, including audio, video, structured, semi-structured, and unstructured data. [Section 1.4, Table 1, Fundamental Function Ingest, page 10]

It is not clear what the term "ingest" actually means. What does it mean to ingest structured data (e.g., an IRIG Chapter 10 data file)? What does it mean to ingest unstructured data (e.g., a Test and Evaluation Master Plan [TEMP] document)? Recommend additional discussion and clarification on what "ingest" means.

3.2.12 Reviewer Comment

REVIEWER NAME: **[REDACTED]**
REVIEWER RANGE: **[REDACTED]**
COMMENT TAG: **CONCERN: COST**

In Section 1.4 the following statement is made:

> Information storage hardware shall be large enough to handle all T&E community data. At first, "all" means all participating programs. As the enterprise is built, "all" will expand to mean all weapons systems test data and all related and relevant integration lab and simulation data. [Section 1.4, Table 1, Fundamental Function Warehouse, page 10]

In order to emphasize that storage should be hardware-agnostic, recommend replacing "information storage hardware" with "information storage capacity".

The cost to store "all" T&E data may exceed what is financially feasible. For example, Section 2.1 states an IOC storage requirement of 100 PB growing to multi-exabytes (EB) of data. Using Amazon AWS GovCloud West S3 cloud storage (cost as of Jan 2020) and an FOC storage requirement of 100 EB, then a first order approximation of the storage cost is as follows:

- IOC: 100 PB = $43 Million/yr (100 PB * 12 mo/yr * 1000 TB/PB * 1000 GB/TB * $0.0355/GB/mo)

- FOC: 100 EB = $43 Billion/yr (100 EB * 1000 PB/EB * $43M/100 PB/yr)

Consequently, recommend changing the first sentence in the STATEMENT above to:

> "Information storage capacity shall be large enough to handle all T&E community data ***as feasible***."

3.2.13 Reviewer Comment

REVIEWER NAME: **[REDACTED]**
REVIEWER RANGE: **[REDACTED]**
COMMENT TAG: **CONCERN: COST**
COMMENT TAG: **CONCERN: STORAGE DURATION**
COMMENT TAG: **CONCERN: DATA WAREHOUSE/DATA LAKE**

In Section 1.4 the following statement is made:

> The enterprise must accurately warehouse data and relevant information on weapon platforms such as aircraft and ships for up to 40 years to ensure there is a background and/or baseline for comparing to unforeseen future issues, block updates, and to equivalent but newer weapon systems. [Section 1.4, Table 1, Fundamental Function Warehouse, page 10]

The term "warehouse" with respect to data storage typically refers to structured data rather than both structured and unstructured data as noted in the following reference:

> Unlike a data warehouse, a data lake is a centralized repository for all data, including structured and unstructured. A data warehouse utilizes a pre-defined schema optimized for

> analytics. In a data lake, the schema is not defined, enabling additional types of analytics like big data analytics, full text search, real-time analytics, and machine learning.[3]

Since the first bullet in the "Warehouse" fundamental function identifies the requirement to store both structured (e.g., test data) and unstructured data (e.g., test plans), recommend replacing the term "warehouse" in Table 1 with the more generic term "store" or "storage" instead.

It is not clear what the "up to 40 years" actually represents. Does this mean store data for "at least 40 years", or does this mean store data for "at most 40 years"? Also, why 40 years instead of 10 years or 20 years or 30 years or the life of the weapon platform or forever?

Recommend rewriting the STATEMENT above to be:

> The enterprise must accurately store data and relevant information on weapon platforms ***for as long as it is feasible*** to ensure there is a background and/or baseline for comparing to unforeseen future issues, block updates, and to equivalent but newer weapon systems.

3.2.14 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: COST

In Section 1.4 the following statement is made:

> The KM/BDA enterprise shall be designed to incentivize stakeholders to use it, not penalize (or charge) them to use it. [Section 1.4, Table 2, Key Non-functional Area Usability, page 11]

Not sure if this requirement is relevant to the KM/BDA enterprise architecture. Who pays for specific instances of the KM/BDA enterprise architecture seems independent of how the KM/BDA enterprise architecture itself is defined. Recommend additional clarification on how this "no usage charge" requirement impacts the architecture.

3.2.15 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: COST

In Section 1.4 the following statement is made:

> Sustainment efforts shall not affect current T&E resources. [Section 1.4, Table 2, Key Non-functional Area Sustainability, page 11]

This requirement seems unobtainable. Given the aggregate infrastructure required to support the KM/BDA enterprise will be significant, it seems that sustainment efforts will most certainly affect current T&E resources. Recommend additional clarification on this requirement.

[3] "Data Warehouse Concepts: How do data warehouses, databases, and data lakes work together?" Amazon Web Services, accessed 23 July 2021, https://aws.amazon.com/data-warehouse/.

3.3 Knowledge Management and Big Data Analytics

3.3.1 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: NONE

The document represents a reasonable strategy and guideline to managing big data requirements on DoD ranges. It covers just about all aspects of development and testing any engineer would welcome in their regular duties.

Here at [REDACTED] range we just don't compile the volume of data as other DoD test ranges, so some aspects of the approach may not rank as high as at other ranges. Nevertheless, the parts of BDA that include "anomaly detection" and "regression analysis" hold value here because of the amount of time we spend certifying range systems and looking for problems.

As a member of the DSG, my vote would be to accept the document as-is and support it as a valuable approach for managing data at ranges.

3.3.2 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: DATA WAREHOUSE/DATA LAKE

In Section 2.0 the following statement is made:

> There are four major functional areas with regard to KM: 1) gathering data into the KM system and associating the appropriate metadata with it; 2) warehousing the data for long periods of time, indexing it, making it available to users to easily search for in a timely fashion when they need it; 3) providing analysis tools and capabilities, such as BDA, that can perform computations on large amounts of data in parallel; yielding products that would normally not be derivable on such large data sets; and 4) provide visualization and reporting, wherein both the vast amounts of gathered data as well as the analysis products can be both visualized intelligibly and reported automatically to the appropriate analyst. [Section 2.0, page 14]

The term "warehouse" with respect to data storage typically refers to structured data rather than both structured and unstructured data as noted in the following reference:

> Unlike a data warehouse, a data lake is a centralized repository for all data, including structured and unstructured. A data warehouse utilizes a pre-defined schema optimized for analytics. In a data lake, the schema is not defined, enabling additional types of analytics like big data analytics, full text search, real-time analytics, and machine learning." (Amazon Web Services, "Data Warehouse Concepts…"

Since the KM/BDA system is required to store both structured (e.g., test data in IRIG Chapter 10 format) and unstructured data (e.g., test plans in PDF format), recommend replacing the term "warehousing" in item 2) with the more generic term "storing" instead.

3.3.3 Reviewer Comment

REVIEWER NAME: [REDACTED]

REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: STORAGE AMOUNT

In Section 2.1 the following statement is made:

> The initial operating capability (IOC) for a T&E KM/BDA enterprise is estimated to be over 100 petabytes, more than a four thousand times larger than Wikipedia, and is expected to grow into the multi-exabyte size range over time. [Section 2.1, page 15]

It is not clear how the IOC estimate of 100 PB was derived. Recommend providing additional information on how this number was determined.

3.4 Software and Data Architecture

3.4.1 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: ROADMAP

I did not have time to read word for word; however, based on a review and prior knowledge I would say in general it looks good. I like that we are looking to reuse information to make better faster decisions. To me the proof will be in the follow-on document that is described in section 3.1. I know there has been success with KM on a large scale ongoing T&E effort, but I am curious as to how someone plans to use seemingly disparate data sets in BDA to generate some useful information. I am sure it can be done, I just think that needs to be well defined and understood before designing a solution for a not yet known problem.

3.4.2 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: HADOOP/STRUCTURED DATA/NOSQL

We're starting to look more into no-SQL databases here, largely due to the nature and form of the data we're storing. Most of the database talk in the document seemed to focus on RDBMS (or row-column, at least) solutions, with only a few passing mentions of no-SQL. I'm not sure if this is because RDBMSs are more common or not, but I wouldn't mind seeing a bit more description of how no-SQL databases would interact with this architecture.

3.4.3 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: HADOOP/STRUCTURED DATA/NOSQL

We are also using a no-SQL database in our current design. We are using Accumulo in our development lab and it will be an integral part to meeting our customers' requirement. I would suggest making sure everyone on this DS-014 task is aware of what Accumulo is. It was developed by the NSA specifically for caveat separation. It has since been turned over to Apache. AFTENCAP is also using Accumulo on their accredited PL-3 system.

3.4.4 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: OTHER ARCHITECTURES/GUIDANCE

In Section 3.0 the following statement is made:

> The BDAS is a software and hardware enterprise with geographically distributed hardware and data locations that will operate for the user as if it were local to them. Not all data and transactions can be local or free of latency, so the distribution is optimized for user needs. In the BDAS, a few large-scale Regional Analytic Capabilities with a large amount of storage and processing power are integrated with smaller-scale installations at local ranges using the JMETC MILS Network (JMN). [Section 3.0, page 17]

According to the "DoD Cloud Strategy" dated December 2018:

> The Department of Defense is driving towards an enterprise cloud environment that is composed of a General Purpose cloud and multiple Fit For Purpose clouds…The Department will implement a commercial General Purpose enterprise-wide cloud solution, Joint Enterprise Defense Infrastructure (JEDI), for the majority of systems and applications. This General Purpose cloud will allow for the Department to take advantage of economies of scale, broadly provide common core services, and ensure information superiority through data aggregation and analysis…The primary implementation bias for DoD will be to utilize General Purpose cloud computing. Only when mission needs cannot be supported by General Purpose will Fit For Purpose alternatives be explored. In such a case, a mission owner will be required to submit for approval an Exception Brief to the Office of the DoD CIO describing the capability and why the General Purpose cloud service does not support their mission.[4]

Will the BDAS Regional Analytic Capabilities centers (RACs) be part of the JEDI General Purpose cloud, or will they be designated as Fit-For-Purpose clouds? Whether the RACs will be part of the JEDI General Purpose cloud or be designated as Fit-For-Purpose clouds, will the JMETC MILS Network (JMN) be an approved interface into these clouds? Recommend additional discussion on how the BDAS will interface with the published DoD Cloud Strategy.

3.4.5 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: NETWORK BANDWIDTH

With off-base connectivity, most likely the user will not "see" the data as being local to them. Unless you shell out millions in large pipes across country, you will never get rid of latency issues. Also, all of this data is classified so encryption also has to be accounted for. Encryption on a small pipe further hurts performance and encryption on a large pipe increases in cost the larger the pipe is if you use Commercial Solutions for Classified (CSfC). I'm assuming since this rides on JMETC they will use KGs which anything above 10Gb is pretty much unobtainable.

[4] Department of Defense. *DoD Cloud Strategy*. December 2018. May be superseded by update. Retrieved 23 July 2021. Available at https://media.defense.gov/2019/Feb/04/2002085866/-1/-1/1/DOD-CLOUD-STRATEGY.PDF.

3.4.6 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: NETWORK BANDWIDTH

In Section 3.0 the following statement is made:

> Data is produced by the ranges and stored locally, either on the ranges' existing storage or on newly deployed storage. Range analysts perform analysis against the locally-stored data, with the support of data scientists, tools, and previously recorded data at the RACs. When local personnel are finished with their analysis of the collected data, and when the data owner agrees, the local data is migrated to the RAC over the network, freeing up local resources for the next test data collection. [Section 3.0, page 17]

The ability to migrate local data to the RAC over the network will likely be a significant problem for many test ranges. For example, given:

1) 1 week = 7 days/week * 24 hours/day * 60 min/hour * 60 sec/hour = 604,800 seconds
2) 1 PB = 1000 TB/PB * 1000 GB/TB * 8 bits/byte = 8,000,000 Gbits
3) 1 Gbps = 1 Gbit/second
4) 94.68% network efficiency = 0.9733 TCP/IP efficiency * 0.9728 Ethernet efficiency[5]
5) network bit rate = network throughput / network efficiency[6]

Then to transport 1 PB of data per week over the network, the minimum network throughput required from the test range to the RAC required would be:

13.23 Gbps network throughput = 1 PB/week * 1 week/604,800 seconds * 8,000,000 Gbits/PB

Hence, the minimum network bit rate required to transport 1 PB of data per week over the network would be:

14 Gbps network bit rate = 13.23 Gbps network throughput / 94.68% network efficiency

Given an imminent capability to collect over 500 TB of data for per test aircraft per test mission coupled with a test range conducting multiple test missions per week and often with multiple aircraft per mission, at least some of the test ranges will likely require multi-100 Gbps network connections between them and the RAC.

Recommend adding an additional section specifically devoted to discussing and emphasizing the importance of network bandwidth in order to adequately connect the test ranges to the RACs as is required to implement the T&E KM/BDA enterprise.

3.4.7 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: NETWORK BANDWIDTH

You can't just take 1 PB of data and say Oh I will send to the cloud.... It might be there by Christmas. A 100 Gb connection to the cloud from each test range is not feasible. I got a quote for

[5] "TCP Over IP Bandwidth Overhead", Packet Pushers, accessed 23 July 2021. https://packetpushers.net/tcp-over-ip-bandwidth-overhead/.
[6] "Ethernet Frame: Maximum Throughput", Wikipedia, last modified 29 May 2021, https://en.wikipedia.org/wiki/Ethernet_frame#Maximum_throughput.

100 Gb link from us to [REDACTED] and it is $10 M a year sustainment cost with ATT/Verizon type carrier. I'm not positive but I don't think DISA can support 100 Gb network connections between bases yet. Infrastructure will most likely have to be upgraded at each test range.

3.4.8 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: COST

In Section 3.1 the following statement is made:

> Most analysts are quite properly focused on answering the specific questions raised in the Test and Evaluation Master Plan (TEMP); however, it is important to ask additional questions so that “unknown unknowns” (problems or issues that the writers of the TEMP did not imagine at the time) can be detected and acted upon in a timely and cost-effective manner. Evaluating all of the data, evaluating one program’s data against another program’s data, and looking for unknown unknowns are the primary motivations for bringing the T&E community into the BDA world. [Section 3.1, page 18]

It is entirely possible that while BDA may reduce the number of tests, BDA may actually increase the overall testing cost. This is because the additional infrastructure costs incurred for the additional storage and computing and network resources needed to enable BDA, along with the additional personnel costs for the additional system administrator and data scientist technical expertise needed to enable BDA, may be greater than the cost reductions realized due to the reduced number of tests.

If implementing BDA may actually increase the overall testing cost, and a primary motivation for implementing BDA is to ask questions outside of the TEMP, then who actually pays for BDA becomes a critical question. If it is the responsibility of the test program to pay for BDA, then the test program manager may be reluctant to pay for the additional cost of BDA in order to answer questions that are outside the TEMP.

Recommend adding an additional section specifically devoted to discussing who actually pays for BDA and how this consequently impacts the implementation of BDA.

3.4.9 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: HADOOP/OTHER BDA FRAMEWORKS

In Section 3.6.2 the following statement is made:

> **Unstructured/Semi-structured Database (Hadoop)** – This element in the architecture framework is the only one in which we have identified a single open source product that will fill the need. That product is Hadoop (hadoop.apache.org). Hadoop is a software framework for both storing and processing large amounts of un-structured or semi-structured data and is the foundation of almost all open source BDA tools. “Unstructured” data is data that has no inherent schema, such as text data; while semi-structured data consists of data that has some structure (like an XML file) but is not necessarily organized into tables of rows and columns (though it can store that type of data as well). Hadoop consists of a replicated distributed file system (called Hadoop File System (HDFS)), a

> software scheduling and resource management package called Yarn, and a parallel processing engine for large-scale data sets called MapReduce. HDFS is the software subsystem for storing all of the data and all of the metadata and schemas associated with the data. [Section 3.6.2, page 26]

Apache Spark (https://spark.apache.org) is another open source product that will fill the need. Apache Spark is a "unified analytics engine for large-scale data processing" and it can be run "using its standalone cluster mode, on EC2, on Hadoop YARN, on Mesos, or on Kubernetes" and can access data in "HDFS, Alluxio, Apache Cassandra, Apache HBase, Apache Hive, and hundreds of other data sources." As specific examples, Spark can be used for simple tasks such as performing word counts and text searches on unstructured data.[7]

Recommend changing the STATEMENT above to acknowledge that multiple open source products can satisfy the Unstructured/Semi-structured Database element in the architecture framework.

3.4.10 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: HADOOP/STRUCTURED DATA/NOSQL

In Section 3.6.2 the following statement is made:

> This element in the architecture framework is the only one in which we have identified a single open source product that will fill the need. That product is Hadoop (hadoop.apache.org). Hadoop is a software framework for both storing and processing large amounts of unstructured or semi-structured data and is the foundation of almost all open source BDA tools. [Section 3.6.2, page 26]

Hadoop can also store and process structured data such as that contained in such as HBase which is a "scalable, distributed database that supports structured data storage for large tables" (https://hadoop.apache.org/). Recommend changing the last sentence of the STATEMENT above to:

> "Hadoop is a software framework for both storing and processing large amounts of unstructured, semi-structured, or structured data and is the foundation of almost all open source BDA tools."

3.4.11 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: HADOOP/STRUCTURED DATA/NOSQL

In Section 3.6.2 the following statement is made:

> While the Hadoop database is the primary store for range data over the long term, it sometimes becomes much more efficient to analyze data in a structured environment. Data can be moved from the unstructured environment to the structured environment and back again depending on the analytic needs of the user. [Section 3.6.2, page 26]

[7] "Apache Spark Examples", Apache Spark, accessed 23 July 2021. https://spark.apache.org/examples.html.

This statement implies that Hadoop is not a structured environment, which is not the case. Hadoop can also store and process structured data such as that contained in such as HBase which is a "scalable, distributed database that supports structured data storage for large tables" (https://hadoop.apache.org/). However, structured data can be stored in traditional relational database management systems (RDBMSs) such as Oracle which require a schema at write and in non-relational (NoSQL) databases such as HBase that require a schema at read. Recommend adding additional discussion and clarification on RDBMS and NoSQL databases, and recommend changing the STATEMENT above to:

> "It sometimes becomes more efficient to analyze data in an RDBMS environment rather than a NoSQL environment. In this case, data can be moved from the NoSQL environment to the RDBMS environment and back again depending on the analytic needs of the user."

3.4.12 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: HADOOP/STRUCTURED DATA/NOSQL

"While the Hadoop database is the primary store for range data over the long term...". I don't think there is anything wrong with that (this is how the NSA currently has their system set up), but we are going to store the long term data in object storage and only move it into the Hadoop/Accumulo environment when a customer needs to do analytics on it. All data in the object storage will have associated metadata tags and caveats so that it can be searched upon and moved to Hadoop for analysis. We went this route because some data generated needs to be stored long term but will never be analyzed. For example we didn't see the need to store a 4-Gb video in Hadoop when that particular customer doesn't have a requirement to do BDA on it. If some other customer needs that video as part of their BDA, it has the metadata tags in object storage and can be transferred to Hadoop.

3.4.13 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: HADOOP/STRUCTURED DATA/NOSQL

In Section 3.6.2 in Figure 7 "Detailed Software Architecture Framework Layered Diagram Overview", the "Unstructured/Semi-Structured Database (Hadoop)" element is identified. However, Hadoop can also store and process structured data such as that contained in such as HBase which is a "scalable, distributed database that supports structured data storage for large tables."[8]

Hence, recommend changing the "Unstructured/Semi-Structured Database (Hadoop)" element to the "Unstructured/Semi-Structured Database" element, and also in the related references in Figure 8 and Table 3.

3.4.14 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]

[8] "Apache Hadoop", Apache Software Foundation, accessed 23 July 2021, https://hadoop.apache.org/.

COMMENT TAG: CONCERN: SOFTWARE DEVELOPMENT

In Section 3.6.2 in Figure 8 "Detailed Software Architecture Framework Layered Diagram", the legend shows the red boxes labeled as "TRMC-Developed Software" and the blue boxes labeled as "COTS/GOTS Software". Assuming that test ranges can also contribute software to the T&E KM/BDA enterprise, recommend changing the legend labels so the red boxes are labeled as "New GOTS Software" and the blue boxes are labeled as "Existing COTS/GOTS Software".

3.4.15 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: HADOOP/STRUCTURED DATA/NOSQL

In Section 3.6.2 in Figure 8 "Detailed Software Architecture Framework Layered Diagram", recommend adding "NoSQL Services" to the "Structured Data Engine" element, and also in the related reference in Table 3.

3.5 Hardware Architecture

3.5.1 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: COST

In Section 4.1 the following statement is made:

> The TRMC's intention is to deploy both processing and storage to each test range to better use the BDA software. [Section 4.1, page 35]

It is entirely possible that while BDA may reduce the number of tests, BDA may actually increase the overall testing cost. This is because the additional infrastructure costs incurred for the additional storage and computing and network resources needed to enable BDA, along with the additional personnel costs for the additional system administrator and data scientist technical expertise needed to enable BDA, may be greater than the cost reductions realized due to the reduced number of tests.

If implementing BDA may actually increase the overall testing cost, and a primary motivation for implementing BDA is to ask questions outside of the TEMP, then who actually pays for BDA becomes a critical question. If it is the responsibility of the test program to pay for BDA, then the test program manager may be reluctant to pay for the additional cost of BDA in order to answer questions that are outside the TEMP.

Recommend adding an additional section specifically devoted to discussing who actually pays for BDA and how this consequently impacts the implementation of BDA.

3.5.2 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: COST

In Section 4.1 the following statement is made:

> The federated nature of the cloud system proposed in this document is the key to achieving scalability, flexibility, and non-interference with existing range operations. New processing, storage, and software is deployed to the ranges in accordance with one of the cases presented above, then all of these systems are tied together in a federated data pool, as shown in Figure 13, below. Given the proper authorization, authentication, and permission, analysts at any range (or with the appropriate network connection and credentials) can see data at any other range or at any RAC, and they can run analysis jobs against all the information at one time if needed. [Section 4.6, page 40]

Given in a big data environment that it is preferable to move the compute to the data rather than to move the data to the compute, it is important to note that if analysts at one range are going to run analysis jobs against the data at another range, then the other range will need to plus up their compute capacity to accommodate these external analysts. If each test range is expected to accommodate analysts from every other test range, then compute capacity plus up at each range could incur a significant increase to the net cost for BDA at each test range.

Recommend additional discussion and clarification on how federating the test ranges together for data access and analysis impacts the BDA infrastructure and cost incurred each test range.

3.6 Security Architecture

3.6.1 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: SECURITY

Given the scope, this should probably be passed along to some of the other RCC groups. In particular, the Cybersecurity Group would probably be interested in the Security/IA portion.

3.6.2 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: SECURITY

Recommend the RCC Cybersecurity Group (CSG) review Section 5.0 of the document.

3.7 Architecture Development & Maintenance

3.7.1 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: RCC/GOVERNANCE

In Section 6.1 the following statement is made:

> The governance vision contains a certain level of centralized management policies and guiding principles that are reviewed and executed by two working groups or entities. A KM/BDA Configuration Management Board (CMB) and Software Development Activity (SDA) will help guide the ranges' data infrastructure to best utilize the KM/BDA enterprise. These two entities, with the support of the TRMC, will lay out the Big Data

> Architecture Improvement and Modernization (I&M) investment roadmap, as well as the blueprint for managing the ranges' big data related policies, procedures, cybersecurity, and resources. Together they will ensure that the ranges' big data infrastructure complies with the DoD CIO Information Technology (IT) management strategy, institutionalization roadmap, and cybersecurity policy guidelines. [Section 6.1, page 48]

There is no reference to the RCC in this section. Is the RCC expected to be a part of the governance strategy? If yes, then how? If not, then why not?

Recommend additional discussion on how the RCC is involved with governance of the T&E KM/BDA enterprise.

3.8 Appendices

3.8.1 Reviewer Comment

REVIEWER NAME: [REDACTED]
REVIEWER RANGE: [REDACTED]
COMMENT TAG: CONCERN: REFERENCES

Recommend adding a "References" section.

4. Recommendations

These recommendations are sorted by comment tag count from highest count to lowest count. Note that comments tagged with "CONCERN: NONE" were not consolidated into recommendations. For a detailed discussion of each recommendation, see the reviewer comments that have a matching comment tag.

As an important point of clarification, these recommendations apply to the architecture framework documented in the TRMC BDAA and do not apply to any specific implementation of the TRMC BDAA.

1. Eliminate cost reductions as a primary benefit of Big Data Analytics seem tenuous.

   COMMENT TAG: CONCERN: COST
   TAG COUNT: 10

2. Correct the apparent misrepresentation of Hadoop and NoSQL databases as not applicable to structured data, and provide additional discussion on usage of NoSQL databases.

   COMMENT TAG: CONCERN: HADOOP/STRUCTURED DATA/NOSQL
   TAG COUNT: 7

3. Explicitly identify and discuss the need for significantly increased network bandwidth between the ranges and the RACs.

   COMMENT TAG: CONCERN: NETWORK BANDWIDTH
   TAG COUNT: 3

4. Have the RCC Cybersecurity Group assess the security architecture.

   COMMENT TAG: CONCERN: SECURITY
   TAG COUNT: 2

5. Identify how the BDAA interacts with other big data architectures and DoD cloud guidance.

   COMMENT TAG: CONCERN: OTHER ARCHITECTURES/GUIDANCE

TAG COUNT: 2

6. Update the T&E KM High-Level Requirements to consistently identify the utilization of commodity hardware and open source software whenever possible.

   COMMENT TAG: CONCERN: COMMODITY HARDWARE/OPEN SOURCE SOFTWARE
   TAG COUNT: 2

7. Replace use of the term "warehouse" with respect to storage with a more appropriate term that applies to both structured and unstructured data.

   COMMENT TAG: CONCERN: DATA WAREHOUSE/DATA LAKE
   TAG COUNT: 2

8. Correct the apparent misrepresentation of Hadoop as the only open source big data analytic framework that can process unstructured and semi-structured data.

   COMMENT TAG: CONCERN: HADOOP/OTHER BDA FRAMEWORKS
   TAG COUNT: 1

9. Clarify how KM and BDA are separate but interconnected concepts, and how BDAA requirements are related to KM requirements.

   COMMENT TAG: CONCERN: KM/BDA/REQUIREMENTS
   TAG COUNT: 1

10. Provide additional clarification on how ranges can "advertise" their private data to other ranges.

    COMMENT TAG: CONCERN: DATA VISIBILITY
    TAG COUNT: 1

11. Show how the IOC storage estimate of 100 PB was derived.

    COMMENT TAG: CONCERN: STORAGE AMOUNT
    TAG COUNT: 1

12. Show how the storage duration of 40 years was determined.

    COMMENT TAG: CONCERN: STORAGE DURATION
    TAG COUNT: 1

13. Clarify how the boundary between TRMC developed software and non-TRMC developed software is determined.

    COMMENT TAG: CONCERN: SOFTWARE DEVELOPMENT
    TAG COUNT: 1

14. Clarify the definition of "ingest".

    COMMENT TAG: CONCERN: INGEST
    TAG COUNT: 1

15. Clarify the definition of "non-traditional data sources".

    COMMENT TAG: CONCERN: DATA SOURCES
    TAG COUNT: 1

16. Provide the RCC with direct participation in the governance of the TRMC BDAA.

    COMMENT TAG: CONCERN: RCC/GOVERNANCE
    TAG COUNT: 1

17. Provide an appendix that identifies relevant references.

    COMMENT TAG: CONCERN: REFERENCES
    TAG COUNT: 1

18. Provide status on the TRMC BDAA Roadmap follow-on document.

    COMMENT TAG: CONCERN: ROADMAP
    TAG COUNT: 1

5. Conclusion

Overall, the TRMC BDAA is a comprehensive document that is well thought out and well written, and provides a good architectural framework for knowledge management and big data analytics within the T&E community. However, there are various concerns with the TRMC BDAA that need to be to be addressed by TRMC before the RCC DSG would be willing to adopt the TRMC BDAA as an RCC standard.

The RCC DSG recommends the following course of action.

1) Revise the TRMC BDAA to address the concerns identified in Section 4. This action would be performed by TRMC.
2) Reassess the revised TRMC BDAA to ensure the concerns identified in Section 4 have been adequately addressed. This action would be performed by the RCC DSG.
3) Recommend whether to adopt or not adopt the revised TRMC BDAA as an RCC standard. As an important point of clarification, this recommendation would only apply to the architecture framework documented in the TRMC BDAA and would not apply to any specific implementation of the TRMC BDAA. This action would be performed by the RCC DSG.

This page intentionally left blank.

APPENDIX A

Citations

"Apache Hadoop", Apache Software Foundation, accessed 23 July 2021, https://hadoop.apache.org/.

"Apache Spark Examples", Apache Spark, accessed 23 July 2021. https://spark.apache.org/examples.html.

"Data Warehouse Concepts: How do data warehouses, databases, and data lakes work together?" Amazon Web Services, accessed 23 July 2021, https://aws.amazon.com/data-warehouse/.

Department of Defense. *DoD Cloud Strategy*. December 2018. May be superseded by update. Retrieved 23 July 2021. Available at https://media.defense.gov/2019/Feb/04/2002085866/-1/-1/1/DOD-CLOUD-STRATEGY.PDF.

"Ethernet Frame: Maximum Throughput", Wikipedia, last modified 29 May 2021, https://en.wikipedia.org/wiki/Ethernet_frame#Maximum_throughput.

National Institute of Standards and Technology. *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*. SP 1500-6. September 2015. Retrieved 23 July 2021. Available at http://dx.doi.org/10.6028/NIST.SP.1500-6.

"TCP Over IP Bandwidth Overhead", Packet Pushers, accessed 23 July 2021. https://packetpushers.net/tcp-over-ip-bandwidth-overhead/.

Test Resource Management Center. *Knowledge Management and Big Data Analytics Architecture Framework*. Version 13. 31 January 2019. May be superseded by update. Retrieved 23 July 2021. Available to people with TRMC credentials at https://www.trmc.osd.mil/wiki/display/JMETC/Big+Data+Knowledge+Management?preview=%2F55968739%2F55968745%2FBigDataArchitecture-v13-2019-01-31-DistA.pdf.

*** * * END OF DOCUMENT * * ***