MTR130279

MITRE TECHNICAL REPORT

Big Data Quality Case Study Preliminary Findings

U.S. Army MEDCOM MODS

Dave Becker
Trish Dunn King
Bill McMullen
Lisa Deifer Lalis
David Bloom
Ali Obaidi
Donna Fickett

September 2013

Sponsor: U.S. Air Force
Contract No.: FA8702-13-C-0001
Project No.: 03137910-AA

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

Approved for Public Release; Distribution Unlimited 13-2568

National Security Engineering Center Bedford, Massachusetts

Abstract

A set of four case studies related to data quality in the context of the management and use of Big Data are being performed and reported separately; these will also be compiled into a summary overview report. The report herein documents one of those four cases studies.

The purpose of this document is to present information about the various data quality issues related to the design, implementation and operation of a specific data initiative, the U.S. Army's Medical Command (MEDCOM) Medical Operational Data System (MODS) project. While MODS is not currently a Big Data initiative, potential future Big Data requirements under consideration (in the areas of geospatial data, document and records data, and textual data) could easily move MODS into the realm of Big Data. Each of these areas has its own data quality issues that must be considered. By better understanding the data quality issues in these Big Data areas of growth, we hope to explore specific differences in the nature and type of Big Data quality problems from what is typically experienced in traditionally sized data sets. This understanding should facilitate the acquisition of the MODS data warehouse though improvements in the requirements and downstream design efforts. It should also enable the crafting of better strategies and tools for profiling, measurement, assessment, and action processing of Big Data Quality problems.

This page intentionally left blank.

Table of Contents

1 Introduction 1-1

1.1 Case Study Organization 1-1

1.2 Participants 1-2

1.3 Case Study Description 1-2

2 Case Study Findings 2-1

3 Conclusions 3-1

3.1 Big Data Scaling Out of the Data Warehouse 3-1

3.2 Types of Big Data Quality Problems 3-3

4 Data Description 4-1

4.1 Characteristics of Big Data 4-1

4.1.1 Kinds of Data 4-1

4.1.2 Volume, Velocity, and Variety 4-1

4.1.3 Meta data 4-2

4.2 MODS Structured Data 4-2

4.3 MODS Geospatial Data 4-3

4.4 MODS Document & Records Data 4-4

4.5 MODS Text Data 4-5

5 Data Operations & Usage 5-1

5.1 Data Usage 5-1

5.2 Operating Environment 5-1

5.3 Data Preparation 5-2

5.4 Data Governance 5-3

6 Data Quality 6-1

6.1 Data Quality Dimensions 6-1

6.2 Special Big Data Quality Issues and Considerations 6-2

6.2.1 Geospatial Data Quality 6-3

6.2.2 Document & Records Data Quality 6-3

6.2.3 Text Data Quality 6-4

7 Data Quality Management 7-1

7.1 MODS Data Quality Management 7-2

7.2 Big Data Quality Tools 7-4

8 References 8-1

Appendix A Acronyms A-1

Appendix B Technology Description B-1

B.1 Data Warehousing B-1

B.1.1 Analytical vs. Transactional Processing Applications B-2

B.2 SSIS Data Quality Services B-3

B.3 Microsoft Big Data Solution [28] B-5

B.3.1 Key Benefits B-5

B.3.2 Broadening Access to Hadoop B-6

Appendix C Project Description C-1

C.1 Background C-1

C.2 System Goals and Objectives C-1

C.3 System Description C-2

C.4 Target Technical Environment C-3

C.5 Conceptual MODS Data Warehouse Architecture C-4

C.6 MODS Data Paradigm C-6

C.7 Known Data Quality Issues and Considerations C-8

List of Figures

Figure 1. Big Data and Data Warehouse Architecture 3-1
Figure 2. MS SQL Server PDW Architecture 5-2
Figure 3. The Architecture of Data Quality 7-1
Figure 4. Generic Enterprise Data Warehouse Architecture B-1
Figure 5. Outline of Microsoft's Big Data Solution B-6
Figure 6. MODS Current Transactional Application Architecture C-2
Figure 7. Depiction of MODS Data Warehouse and Data Mart Layers C-3
Figure 8. Conceptual MODS DW Architecture C-4
Figure 9. MODS Data Paradigm C-6
Figure 10. Conceptual Model Example C-7

List of Tables

Table 1. MEDCOM MODS Data Quality Issues C-8

1 Introduction

Big Data applies to data sets of extreme size (e.g., terabytes, petabytes, exabytes, and zettabytes) that are beyond the ability of manual techniques and commonly used software tools to capture, manage, and process within a tolerable timeframe.[1] The assumption underlying this case study is that Big Data will be prone to the same quality problems that plague traditionally-sized data sets, characterized by accuracy, precision, completeness, consistency, timeliness, lineage, and relevance. However, these quality dimensions are not yet well understood for Big Data, and may be very different from the quality dimensions for traditional sized data sets; i.e., may require completely different strategies and tools for profiling, measuring, assessing, and action (possibly, corrective, informative, or other) processing.

The intent of this MITRE Mission Oriented Investigation and Experimentation (MOIE) project is to conduct a series of Big Data Quality case studies for a diverse set of Big Data initiatives or projects. The case studies reflect literature surveys, documentation reviews, and in-depth interviews with initiative experts and domain subject matter experts (SMEs). Occasionally, the analysis undertakes some exploratory data profiling, measurement, and assessment regarding Big Data Quality issues, tools, and techniques. These case studies capture and document data quality problems and processing techniques that are primarily characteristic of Big Data. The case studies also provide a foundation from which to propose a new Big Data Quality Framework (evolved from prior MOIE data quality research) for use in data management of current and future Big Data initiatives.

This report is one of a set of four case studies documented separately. A summary overview report will also be prepared. This document covers a particular case study that involves a new data warehouse project called Medical Operational Data System (MODS) for the U.S. Army's Medical Command (MEDCOM), the first phase of which involves medical readiness. Subsequent phases will deal with health related human resources, behavioral health, the Public Health Command, Patient Centered Medical Home and other areas.

1.1 Case Study Organization

Each of the four individual case studies is generally organized as follows: Section 1 introduces the full MOIE project and the case study itself. It includes a short description of the context of the study including the general situation and technology involved and the particular project or initiative which is using or generating the Big Data.

For readers interested in additional background: Appendix B includes more detail on the technology involved. Appendix C includes or references a more detailed description of the project using the technology. In order to achieve more impactful readability, the findings and conclusions are presented immediately following the introduction in Section 1. Section 2 distills the information from the rest of the study into a list of key findings that summarize the takeaways related to Big Data Quality in this area. Section 3 summarizes and states conclusions to be drawn from the study and possible directions for future work. The remaining sections provide the meat of the case study analysis, discussing the data, its use, quality problems with the data, and management of those quality problems. Section 4 describes the data with particular emphasis on the kinds of data in the project and its Big Data characteristics. Section 5 describes the practical aspects of how the data is used, the systems/applications involved, their operating

[1] Jeff Butler, "Big Data and Information Quality in the IRS Research Community," *2012 MIT CDO and Information Quality Symposium*, Massachusetts Institute of Technology, Cambridge, MA, July 17-19, 2012, which states that "roughly 80% of all data warehouses in the U.S. are still less than 20TB, the current threshold for "big data."

environments, how the data is prepared and processed, and how the data is managed/governed. Given the background understanding of the data established in Sections 4 and 5, Section 6 begins to look at quality issues that have arisen within the context of this particular Big Data case study. Section 7 then explores how those Big Data quality issues identified in Section 6 are managed, including any special tools and techniques used.

1.2 Participants

Participants in the case study included MITRE employees associated with the MEDCOM project: Lisa Lalis (ldeifer@mitre.org), David Bloom (dbloom@mitre.org), Dr. Ali Obaidi (ali@mitre.org), and Donna Fickett (dfickett@mitre.org). The case study also included members of the MOIE team: Dave Becker (dbecker@mitre.org), Trish Dunn King (dunnp@mitre.org), and Bill McMullen (mcmullen@mitre.org). The MITRE employees associated with the MEDCOM project coordinated interaction with various MEDCOM participants and stakeholders, some of whom were civilian and military personnel, and others who were employees of MEDCOM contractors. The participants collaborated for this paper primarily using scheduled meetings and workshops, teleconferences, email, and phone conversations.

1.3 Case Study Description

While other case studies in this overall investigation deal with initiatives that are already considered to be Big Data in nature, either through the size of their data sets or the technology being employed, the intent of the MODS case study is to explore a typical application that **might** become a Big Data initiative in the future. This might occur through the addition of requirements that expand the volume, velocity, variety, or variability of the data with which the application must deal. In this case, the application we are looking at is the MODS data warehouse. This study will explore the potential changes the data warehouse may encounter in the above areas.

Today a common type of application environment that has approached the scale of Big Data has been the data warehouse. A standard data warehouse consists of a large assemblage of data sets from various sources collected together for the purpose of providing a central place for the normalization or dimensionalization of the data so that it can provide a single source of truth to support the generation of enterprise reports and provide the organization with business intelligence analysis. Specialized software called ETL (Extract, Transform & Load) is used to move data from its source into the data warehouse. Specialized software and hardware is then used to provide high speed relational query access to the data to support various analytics, data mining, presentation, and dashboarding applications. Data Warehouses have been extremely successful at providing organizations with enterprise level capabilities allowing them to view summaries of the full extent of their operations, highlighting topics of interest, and exploring more deeply those aspects that require serious attention.

Data quality (DQ) problems with the data sets that comprise the data warehouse are frequently uncovered during the ETL process when the data is loaded or updated. DQ problems can also be found during testing of the business analytics and intelligence applications that are run against the data, as well as by users after the warehouse and its data have been put into production. Because the data warehouse typically does not “own” the data, it should not change the data (cleanse it). This is not to be confused with data standardization and transformation activities which are necessary for ETL processing. Instead the data warehouse must serve the role of a good steward, and provide feedback through notification of data quality problems to the authoritative sources of the data. Those authoritative sources can then properly address the quality problems in the data for which they are responsible. The data warehouse should also make any data quality problems or quality levels known to the data warehouse users of the data, so that they can factor this into their analysis and decision making.

The question is, what happens when these environments are asked to dramatically scale out (i.e., acquire, store and process) to support new types of data (Big Data) that the organization may wish to use? What types of Big Data will confront them? How will they architect the solution? Are there different types of quality problems they think will have to be addressed?

The MEDCOM MODS project illustrates this situation. Until now, MODS have primarily focused on providing medical readiness Online Transaction Processing (OLTP) capabilities to facilitate the overall medical readiness of the Army forces. The new effort intends to add an Online Analytical Processing (OLAP) to the existing OLTP capabilities. This will be accomplished by creating a data warehouse that is being architected for current expected levels of growth. Size wise, the new data warehouse will be in the tens of terabytes range. For a data warehouse, this will be large, but not exceptionally large. However, if requests for other data sources (such as geospatial data, clinical textual materials, or document and records data) are added to its requirements, MODS could potentially experience explosive growth at some point in the future moving it into the Big Data arena.

This case study explores the data quality aspects of such growth. A more detailed description of the data warehouse technology employed is found in Appendix B. A more detailed description of the MODS project itself is in Appendix C. The following sections describe the data, data operations, and data usage involved in the project, as well as the data quality and data quality management aspects of the project.

2 Case Study Findings

This section presents findings about data quality for MEDCOM MODS Big Data.

Finding 1: Data warehouses should not cleanse data. (*Reference Section 7.1)*

Because the data warehouse does not typically "own" the data, it should not change the data (a narrow definition of the phrase "to cleanse it"). As part of an ETL process, it is almost always necessary to standardize and transform the data so that it can be loaded into and processed by the data warehouse appliance and Business Intelligence (BI) tools. But errors in the values of the data should never be modified. Instead the data warehouse should serve the role of a good steward, and provide feedback to the authoritative data sources of any data quality problems encountered. The data owner would then be responsible for properly addressing the problem. If there are problems where the data cannot be corrected at the source, then appropriate work-arounds adopted by the data warehouse must be documented and communicated to the data owners. The data warehouse should also make any data quality problems or quality levels known to the data warehouse users of the data, so that they can know its reliability and factor this into their analysis and decision making.

Finding 2: Automated device-capture of geospatial data improves all dimensions of data quality. (*Reference Section 6.2.1)*

The adoption of automated device capture of geospatial data for MODS will actually dramatically improve its quality while at the same time making it much easier and more efficient by removing manual entry steps and eliminating human generated data entry errors.

Finding 3: The precision of geospatial data must be matched to privacy and security requirements. (*Reference Section 6.2.1)*

The precision of geospatial data offers great opportunities for analyzing and improving the operations of the MEDCOM community. However, it also creates a new vulnerability which must be closely guarded. Detailed information regarding the exact location and movement of equipment and personnel must be tightly matched to the privacy and security requirements of the data, and access and security levels of the users and applications must be checked.

Finding 4: Meta data creation & management must be automated as much as possible. (*Reference Section 6.2.2)*

Meta data (descriptive information about the data) is extremely important in Big Data management, especially in document and records management. Meta data (e.g., owner, source, creation date, last update, derivation algorithm, etc.) allows the data to be productively used and properly managed. Actually the quality of much of the meta data for MODS document and records data is currently acceptable given the way it is manually verified through many steps. However, the potential future problem is just that – that much of the meta data in MODS is currently created and managed manually. For future Big Data scenarios, it will no longer be feasible for meta data to be entered and updated manually. It must be automated as much as possible.

Finding 5: Current data curation schemes will not work at scale. (*Reference Section 6.2.2)*

Data curation involves the manual discovery of data sources, cleansing and transforming the new data, semantically integrating it with other local data sources, and deduplicating the resulting composite. Current approaches to data curation are very labor intensive [2]. The ability to address data quality problems that data curation attempts to resolve can only be achieved for Big Data sets by adopting a technical program that adds significant levels of automation, advances in machine learning, application of statistical techniques, and/or adoption of new crowd sourcing approaches to human involvement, as long as they don't conflict with the need for privacy/security.

Finding 6: Legal and compliance requirements associated with document and records data will not go away. (*Reference Section 6.2.2)*

A large part of the reason for formal retention strategies and document management policies is to ensure that the creators, managers and users of the documents and records do not run afoul of various rules and regulations, or become ensnarled in serious legal entanglements with the incumbent penalties and problems these situations entail. This can minimize the amount of operational risk assumed by the organization and its people. The amount of regulatory compliance and the level of societal litigation do not appear to be abating any time in the near future, despite the onset of Big Data.

Finding 7: If information is in free text, then retrieval at scale will depend on the state of the art automated indexing and extraction capabilities at scale, which are now emerging and evolving. (*Reference Section 6.2.3)*

Textual data is highly complex compared to structured or semi-structured data, because of its reliance on higher level contextual information, including document structure (e.g., headers, tables, check lists), cross sentence and cross-paragraph references and use of jargon and telegraphic forms. Because of this complexity, standard cleansing methods cannot be readily applied; for example, standard spelling correction is problematic for text where there is heavy use of abbreviations. For Big Data collections of text, standard data cleansing becomes impractical. In many cases, the best that might be expected is to accept whatever quality levels are present, and just deal with the level of uncertainty involved. If specific passages or manageable subsets of importance can be isolated for which high quality is critical, then limited cleansing efforts using traditional techniques could be applied. Interestingly, some text extraction techniques, that are assumed will be used in any future MODS text usage scenarios, and can actually improve extraction performance without requiring expensive cleansing efforts when they are supplied with larger amounts of data [29]. However, this will need to be verified for each usage scenario, and a prototyping strategy would be highly recommended both to establish the cost and value involved in any textual data approach being considered and to select an appropriate way forward.

Finding 8: Transactional users generally have higher quality requirements than analytical users. (*Reference Section 7.1)*

Different categories of users typically have different quality requirements. In general, the quality requirements of transactional users who are trying to get work done are higher than those of analytical users who are more interested in higher level data aggregation in order to report trends and patterns. Analytical users can tolerate lower levels of quality than transactional users. Analytical users can piggyback on quality improvement efforts of transactional users. The implication of this is that an implementation strategy of satisfying tactical transactional users first and then analytical users might be more successful from a data quality perspective, than an implementation strategy that emphasizes satisfying analytic users first and then transactional users.

3 Conclusions

The basic questions for this case study are: 1) what will happen when the MODS data warehouse environment is asked to dramatically scale out to support new types of data (Big Data), and 2) are there different types of data quality problems they think they will have to address?

3.1 Big Data Scaling Out of the Data Warehouse

The approach taken by MODS is to establish a dependence on its selected data warehouse provider (Microsoft) with the expectation that the Microsoft technology will be able to scale up to handle the hundreds of terabytes of data and thousands of users that they could potentially encounter. In addition, Microsoft technologies will provide an environment to properly analyze, extract, and process the data for potential business value. This approach seems reasonable given that all of the major data warehouse vendors (Microsoft, Teradata, Oracle, IBM, etc.) are rushing to fill this Big Data hole in their product strategies.

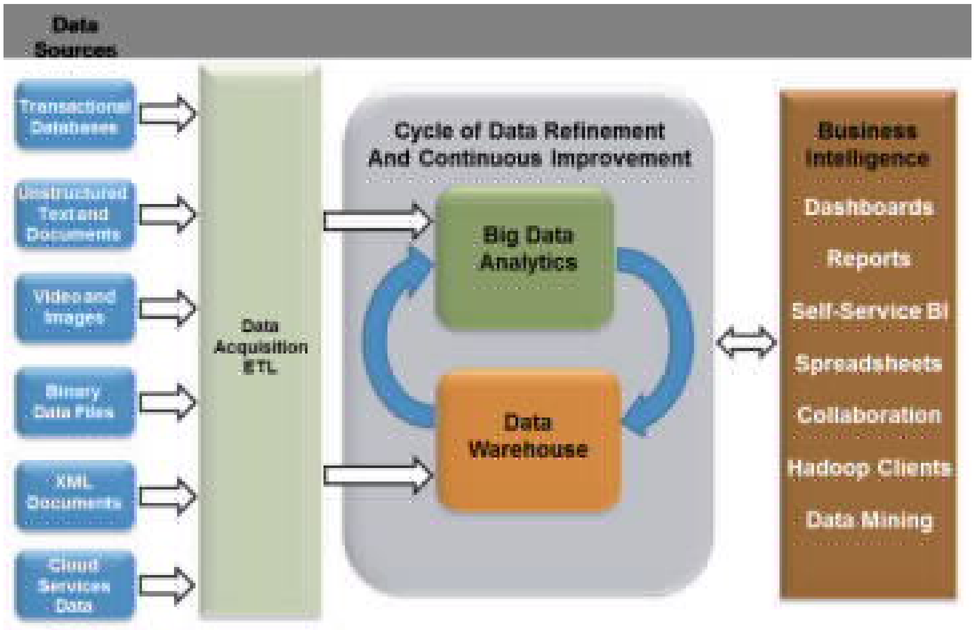


Figure 1. Big Data and Data Warehouse Architecture

Most likely, MODS will end up adopting the typical high level architecture being used for many combinations of Big Data and Data Warehousing as depicted by Johnson & Zahavi [3] in Figure 1 above. It consists of a two-part strategy. Part One provides a Big Data integration hub (here called Big Data Analytics) that allows large amounts of data to be rapidly added to the Big Data repository regardless of its format or quality. Part Two is the traditional data warehouse which is typically reserved for highly normalized or dimensionalized, high-quality, structured data used to drive traditional business intelligence capabilities. However, a continuous improvement loop would cycle between the two parts.

Another component could be added to the Johnson & Zahavi model that addresses Master Data Management (MDM). ISO 8000-110 defines master data as data held by an organization that describes the entities that are both independent and fundamental for that organization, and that it needs to reference in order to perform its transactions [4]. Good MDM has proven critical to effective data quality management. The MODS data warehouse project has indicated a desire to establish an MDM program to facilitate both the management of the quality of the data needed for transaction processing of the data warehouse feeder systems, as well as the reporting and analytics in the data warehouse. The MDM catalogues established and used in the transactional systems and the data warehouse can also be productively employed in the Big Data integration hub and the Big Data analytics environment. The Big Data integration hub would be where large volumes of unstructured and semi-structured data would be loaded and where the Big Data analytics of various types would be performed. This is "uncurated" information, meaning that it has little or no structure (data model) in effect, and has not been reviewed for usefulness or various data quality dimensions [5]. As particular pieces of data are analyzed in the Big Data cluster, and found to be of particular value, they can be transformed, cleansed, and normalized or dimensionalized (the process called data curation), and fed into the data warehouse for more targeted and high-speed traditional analysis by end users.

Likewise, as knowledge and refined data are uncovered in the data warehouse, they can be fed back into the Big Data platform for historical storage as well as for utilization by the more broad brush, time-sensitive discovery queries and analytics that occur there.

Interestingly, this vendor-based Big Data strategy is the same approach which is being explored by the Global Combat Support System-Air Force (GCSS-AF) Data Services group in scaling out their current Teradata-based data warehouse. They plan on using the existing Teradata relational technology in concert with their newer intermediate Aster products and the standard Hadoop map-reduce Big Data technologies. While the single vendor-based strategy has obvious benefits, there can also be drawbacks. For example, the Microsoft SQLServer tools, while reportedly well-integrated, still require a significant amount of custom coding. This single vendor-based strategy where the architectural components have already been integrated, can be contrasted with a best-of-breed strategy where the customer itself attempts to assemble and integrate the best available components, implementing only desired functionality, and achieving a higher level of performance, but also assumedly requiring a not insubstantial time and cost. This debate has been raging for quite some time, and there are good arguments either way [31, 32, 33]. The point is that these decisions require a careful evaluation of the different tradeoffs. MODS has decided to pursue the single vendor approach more fully described in Sections 5.2 and 5.3, and Appendices B.2 and B.3.

> For MODS, the most likely scenario is as follows. As the structured data collection grows, the new data will simply be added to the existing data warehouse, where it will more fully utilize the Massively Parallel Processing (MPP) analytical capabilities of the Microsoft Parallel Data Warehouse (PDW). The Geospatial data will be integrated with and used to augment the structured data already in the data warehouse. The document and records data as well as the text data will be processed through a separate Hadoop-based Big Data Analytics environment using the PolyBase Big Data integration component. This environment would be connected to the data warehouse by various key fields for augmented in-depth information analysis. It is also reasonable to expect that as new data sources are evaluated for inclusion in MODS, they may be loaded first into the Big Data Analytics environment where they can be analyzed in relation to the rest of the overall collection, and as appropriate, curated and migrated to the data warehouse environment, or retained in the Big Data Analytics environment, or discarded altogether.

3.2 Types of Big Data Quality Problems

The current proposed architecture for MODS will probably be able to handle current projected growth into the tens and possibly even hundreds of terabytes. So, MODS will probably not encounter Big Data in the growth of their existing structured data collections and sources. MODS will most likely encounter Big Data when they start to add geospatial data, document and records data, and text data.

MODS has already identified a number of quality issues (and associated data quality factors) in the structured data from the current list of their required source systems (see Table 1 in Appendix C). While additional efforts to improve identified quality issues can be undertaken and will have significant benefits, they are not the primary focus of this study.

Oddly enough, any quality problems MODS has today with geospatial data will probably improve if the choice is made to expand coverage in this area. This is because the addition of geospatial data at finer precision levels will have to be generated by instrumentation rather than by hand as is currently in use. Today's geospatial instrumentation generates geospatial data in much greater volumes and at much higher levels of accuracy than manual-annotation. The area that is challenging and needs focus is in the matching of the privacy and security requirements of the customers to the levels of precision that can be made available through the instrumentation technology available. Unless this data is extremely well managed, the expansion of geospatial data carries the possibility of putting very sensitive information into the hands of unauthorized individuals.

Document and records data could be implemented in a number of ways. There have been very good non-Big Data technologies available for quite some time for managing very large collections of documents and records that have been computerized through scanning and OCR; for example, IBM's FileNet Content Management. While it can be considered a reasonable approach to put into place, these systems are frequently expensive to implement, complex to administer, and often have scale limitations of their own.

Where document and records data can also become interesting, is when the textual content of these documents and records is subjected to various analytical techniques to extract information or characterize the content in different ways (also known as data or text mining). This situation is analogous to the additional free text material MODS is considering adding from its various existing structured and semi-structured sources. Due to the complexity of handling textual data, the measurement and management of the quality of these materials is extremely difficult.

Where quality is critical, an involved process, typically with humans in the loop, must be engaged to identify and resolve any issues of accuracy, lineage, or consistency. Part of this process is curation (mostly related to data and meta data management), and part may be cleansing of textual data errors (which would be required to be performed by the authoritative source). MODS has evolved a good process for the curation component. For example, some of the meta data in various health profiles available through MODS will often be of very high quality because the meta data is reviewed in each step of a long, involved, manual workflow.

To date, MODS has essentially ignored textual data correction, choosing to leave this data alone and use it only for reference and corroboration. If the future intent is to add a requirement to include dramatically more textual data, then the system must be able to handle it. Further, if the requirement is to begin automatically extracting information from this data, then MODS will have to tap into new, scalable solutions that are beginning to emerge that can reliably extract information from textual data without extensive cleansing. However, if the requirement will also demand very high quality levels in the textual data comparable to that found in structured data, then MODS may need to work with the authoritative systems owners to ensure they assume responsibility for performing any needed cleansing activities to make the textual data usable, and to find viable, scalable alternatives (see discussion in Section 4.5 below).

The typical response in many Big Data collections of textual data is simply to accept that there will be different degrees or levels of quality [34]. Any analytical processing performed on this data is generally geared toward general exploration and discovery or background investigation. Specific results uncovered in these collections are assumed to have a quality level associated with prior experience with the data sources, and/or an externally established source data reputation made available from an authoritative agent. Answers returned from this type of collection that will be used in critical applications must often be double-checked or corroborated through alternative sources of the information, thus adding extra cost and time to their appropriate use. MODS will have to be very careful as they expand into this area by making sure that the use of the data is properly synchronized with quality levels needed by its applications.

4 Data Description

This section provides an overview of data as it is typically characterized in a Big Data project. It then explores these characteristics as they apply to the different kinds of data in the MODS project.

4.1 Characteristics of Big Data

This section discusses three important data aspects of a Big Data initiative (kinds of data, the 3 V's, and Meta data).

4.1.1 Kinds of Data

There can be many different kinds of Big Data. One obvious way to differentiate them is to classify them as Structured Data, Unstructured Data, and Semi-Structured Data. Structured Data has a defined length and format, and includes row and column, fielded, relational tabular data usually generated as byproducts of doing a transaction or operation, or as the output of a sensor. Unstructured Data has little or no repeating format information from one instance to the next and includes textual, audio, video, image, radar, scientific, social media, or website content data. Semi-Structured Data is data which is unstructured with some repeating format information that allows for specialized processing; for example, XML or EDI data [6]. The important point is to establish a basis for distinguishing between the kinds of data while still allowing for comparison and contrast of different collections of Big Data.

A further distinction which is sometimes useful for analysis is whether the data is Human-generated, or whether the data is Computer or Machine-generated. Human-generated data is data that humans, in interaction with computers, supply. Computer-generated or Machine-generated data is data that is created by a machine without human intervention [7].

The data used by the MODS project is currently mostly structured data fed from a number of different database sources. In the future, MODS might entertain new requirements that would introduce large volumes of other highly structured data, such as geospatial data, as well as significant amounts of unstructured and semi-structured material to include document and records data, as well as clinical textual data. It is these potential future data growth requirements driving large expansions of data coverage that could move MODS into the realm of Big Data.

Furthermore, there are other kinds of data that have not yet been considered as part of this case study that could be included in future data growth requirements. These include clinical multimedia records such as x-rays, CDs, pdf documents, word documents, videos, photographs, power point presentations, social media posts, emails, customer service calls, etc.

4.1.2 Volume, Velocity, and Variety

Big Data collections are often described in terms of "3 V's:" Volume, Velocity, and Variety. These terms are used to describe the complexity associated with using a data collection.

- "Volume" characterizes the size and growth rate of the data collection.
- "Velocity" considers the frequency of data generation or data delivery for the data collection. Velocity can also consider how quickly the data needs to be analyzed and how fast can become obsolete.
- "Variety" describes the number of different sources , logical entities, data types and formats

that are contained in the data collection. For MODS the data sources are internal or external systems or databases that will provide the data represented by each data category . These categories represent the information needed to answer the basic business questions to be addressed to the system.

We will discuss these dimensions for each kind of data that MODS will deal with or might deal with in the future.

4.1.3 Meta data

A key feature of all data is what is called meta data, commonly referred to as "data about data." Meta data can be defined as any descriptive information about the data that allows it to be productively used and properly managed. Meta data reflects data transactions, data events, data objects, and data relationships. Meta data can be classified as descriptive meta data, structural meta data or administrative meta data [8]. Meta data management is the set of processes that ensure the proper creation, storage, integration and control to support associated usage of meta data.

Meta data regarding dates, times, locations and identifiers of the medical readiness data in MODS is captured throughout the process on the forms and within the profiles associated with the events and incidents. This data is used for filing and retrieval of the data, and for general management of the information.

4.2 MODS Structured Data

MODS structured data includes row and column or table data that can be categorized as both transactional and reference information where the transactions include shorthand codes or references to the more fully described reference or master data. Transaction-oriented data is considered to be naturally time-variant and reflects tracking information about activities performed. Reference data is descriptive data that generally characterizes those activities.

The structured data covers a full range of needed deployment information including personnel descriptions and their assignments, deployment status, limiting deployment statuses, mental and physical health profiles and readiness assessments, medical referrals and appointments, and screenings, allergy and other medical warning tags, blast and hazardous material exposures, immunization status, drug prescriptions, DNA tracking, etc. Currently MODS does not receive highly sensitive deployment information, it could in the future receive historical deployment information from post-deployment health assessments.

Note that some of the entities hold textual information such as medical evaluations.

Volume - The structured data that is currently planned for loading into MODS constitutes approximately 16 terabytes. There are plans for accommodating natural growth and extension of this collection to approximately 60-80 terabytes. While this is large, it is not exceptional among data warehouse implementations. The important aspect of it is the growth rate as it adversely affects the handling of the data quality issues.

Velocity - The MODS data collection will be populated from its various feeder operations systems. These interfaces will run on a daily, weekly, and monthly basis. This does not constitute a significant velocity problem. However, data analysis requirements are expanding, and the MODS data warehouse is providing new hosts of Business Intelligence reports that will require the data to be available on time. The Army is moving to agile deployment and needs to be able to provide deployments within a very short timeframe to address global geopolitical challenges.

Variety - The MODS data collection will be populated from various operational systems. While MODS is evaluating receipt of data from over 25 different possible data sources, in actuality it will probably focus most attention on only a few of these. For a major data warehouse project, this is not an overly large number of sources. MODS has identified over 40 different data categories which range from personal and organizational identification information to general disease and treatment protocols, to individual medical profiles, to medical staff allocations. Each of these categories will include multiple fields that detail the structure of the information. Each of the categories can be supplied from one or more source systems, each of which has its own formatting conventions. While conversion and reconciliation of this variability in the data will constitute a major portion of the MODS implementation work, this degree of complexity seems to be consistent with other Big Data Warehouse implementations.

4.3 MODS Geospatial Data

Geospatial data are defined as "information that identifies the TRACKING geographic location and characteristics of natural or constructed features and boundaries on the Earth" [9]. Geospatial data originally derives from the automation of paper maps that included 'gridding' and key coding of map-based attributes providing layers of information [10].

Because geospatial data has a very complex and repetitive structure, it could be considered structured data. Also, since it rarely, if ever, exists separate from the content to which it applies, it should not be distinguished separately from that content [11].

Geospatial data is naturally very important for the Army, and there are many projects deeply involved with its collection and application. Mobility today means every person or piece of equipment is potentially a sensor gathering geospatial data. Wherever we go, whether we walk, drive, bike or fly, we are continuously producing new location-based geospatial data. Visualization and analysis of that data can make it easier for us to make everyday decisions.

For MODS, this geospatial data would be associated with the location and movement of troops, environmental sample mapping, injury pattern, multi-dimensional representations of events, movement of the medical personnel, resources, and equipment needed to support the troops [12].

Volume - The current MODS geospatial data is very minor and coarse grained. It basically consists of identification of the military facilities and locales where the soldiers and medical personnel are based and operating, and includes almost no latitude/longitude (lat/long) information.

However, given the capability of modern mobile devices, this small amount of location information could be exponentially increased. The amount of geospatial data in a given collection will depend on the number of entities that you want to be able to track, how frequently you want to collect the location data, and what level of resolution is needed. Geospatial data presents a whole new level of processing. For example, every time the resolution of an image or a raster elevation data set is increased two-fold, the size of the data quadruples. Geospatial data could quickly become very voluminous [13].

Velocity - The current MODS geospatial data is collected only on an event-driven basis associated with the providing of health care services or scheduled check-ups or shots, or for the treatment of injuries or illnesses. However, in the future it could be generated, collected, and processed on a near real-time basis, and in great detail in geographically dispersed locations.

Variety - The current MODS geospatial data consists of identification of the military facilities and locales where the soldiers and medical personnel are based and operating. This data is manually entered into the fields in written and electronic reports and profiles. In the future, this approach would be completely transformed to utilize Global Positioning Satellite (GPS) devices and Geographic Information Systems (GIS) systems to provide geospatial data input. There will be many such devices and systems which will be generating this data. Instead of textual or coded forms and database fields, this data will be collected in the standard formats for various GPS devices and GIS systems. This is not high in complexity, but is high in volume.

4.4 MODS Document & Records Data

Document and records data refers to the significant business documents of the organization. They can exist in physical or electronic form. These materials have significance because they provide evidence of the organization's business activities, and are frequently used to satisfy regulatory compliance requirements. As such, they need to be carefully managed through their entire lifecycle [8].

Across the federal government, huge collections of documents and records can be found in many different forms (paper, digital, and multimedia) [14]. U.S. Army MEDCOM is no exception to this situation. Older health profiles exist in paper form, and will be scanned and processed through OCR software. Newer health records are created electronically as a combination of textual information and medically coded data entered by hand at the time the health profile is generated, or a health event occurs and is recorded.

These documents and records are primarily needed for various clinical and readiness substantiation purposes. Also, there are various mandates driving the capture, management, and retention of this information, including maintenance of the security and privacy of sensitive information; for example, Health Insurance Portability and Accountability Act (HIPAA) compliance and Personally Identifiable Information (PII) designations.

Volume - Currently the volume of this type of data scanned and captured in electronic form is relatively small, on the order of hundreds of thousands of health profiles at say around 500K per profile, or in the hundreds of gigabytes. There has been resistance up until now in gathering and/or expanding the current small document store because of the need to control the scope to the current mission and support for deployment readiness. However, given the stand-up of the MODS data warehouse, and its implementation of better opportunities for analytics and data mining, it is anticipated that the mission will expand dramatically in the next two-three years. Should the mission be expanded, the number of records could increase 100 fold, and the inclusion of additional business areas could easily move the document store into the multiple terabytes range.

Velocity - The generation of this data would still be driven by event or incident, but there could be a significant increase in the number and precision of recordable events.

Variety - Today, information about the recordable events that are captured in various documents and records is entered manually by over 20,000 clerks located at medical facilities all over the world. With the adoption of new equipment and systems, points of data capture could be multiplied many fold. There could also be a corresponding increase in the number of different data structures and formats to be accommodated.

4.5 MODS Text Data

There is perceived to be a lot of future potential in acquiring and analyzing unstructured free text information that currently exists in doctor notes, clinical notes, nursing notes, and various case management documents. Unstructured textual data is generated to be consumed by a human reader; it is very efficient at conveying complex nuanced information – to another person. It eschews the need for complex hard-to-maintain coding schemes needed to capture structured data.

However, these properties make it very different from structured data, and the two have different requirements, data quality issues, and support different use cases. Unstructured (textual) data is subject to a very different set of data quality measures than structured data – it often makes little sense to put them side by side and talk about textual data as "errorful and ambiguous" and therefore needing to be cleansed or discarded outside of a specific usage context. It is important to understand what kinds of information are contained in textual data, and what needs to be extracted from the data (as is), and how good current methods are for extracting data for human consumption and further automated processing and aggregation.

Should this textual information be captured in electronic form, then different language identification, handwriting recognition, and Natural Language Processing (NLP) techniques could be used to extract information for subsequent analysis. Today this information is strictly for reference purposes only; it is not used in automated analysis.

Regarding automated analysis, "[o]ne way to tap into the potential of unstructured data is through text analytics. Text analytics is the practice of semi-automatically aggregating and exploring textual data to obtain new insights by combining technology, industry knowledge, and practices that drive business outcomes. ...Combined with the analysis of structured data, text analytics can help businesses in their efforts to uncover signals and patterns" [15].

Volume - The free text fields vary in length from say ten characters to several hundred characters. Other materials can be much more extensive, and are frequently made available as referenced material or attachments. It might also be possible to begin including audio or video of the incidents or events. If the scope of MODS should expand to include this type of material, the increase in volume could again involve exponential growth, easily within the tens or hundreds of terabytes.

Velocity - Today, the text associated with these recordable events are entered manually by administrative and medical personnel located at medical facilities all over the world. With the adoption of new equipment and systems, points of data capture could be multiplied many-fold.

Variety - Sources for MODS textual data might be Armed Forces Health Longitudinal Technology Application (AHLTA) and the Military Health System (MHS), both of which capture a lot of their data in text fields. Textual material is extremely complex; it is currently challenging to extract useful information at scale from text. For controlled fields on forms, the success rates are much higher.

5 Data Operations & Usage

This section first describes the different uses/applications that are intended for the data. Next it describes the physical operating environment and software in which the applications execute. It then discusses how the data is processed and prepared for use. Finally, it discusses how the data is managed/governed as an enterprise asset.

5.1 Data Usage

This section describes the uses of Big Data. It provides information about the overall objectives of the initiative, the specific application(s)/tool(s) that will be employed to meet those objectives, and how the data will be employed by those applications. This can sometimes be represented in a set of use cases that detail the processes through which the data will be accessed.

MODS stakeholders initially identified over fifteen Business Improvement Opportunities (BIOs) primarily related to three Levels of Efforts (LOEs) that better identify, provide care for, and reduce the Army's Medical Not Ready (MNR) soldier population. The MODS stakeholders then defined over 100 different Business Questions (BQs), which represent analyses required to support, enable, or otherwise measure the defined BIOs. This analysis effort essentially established the baseline for the requirements of the MODS project. The structured data aggregated into the MODS system is intended to be able to provide accurate, timely, and comprehensive answers to these BQs.

Assuming the quality of the additional data were adequate, the new data could be put to many uses. Geospatial data (long/lat, zip, state, etc.) would be used to more precisely identify where incidents have occurred, to track where processing of sick and injured takes place, to identify who is being treated, to follow their movement through the system, to evaluate the effectiveness of treatment programs and protocols, and to audit the workloads and staffing plans of medical personnel. It can also identify the locations of environment changes that affect the soldiers and finally, identify blast locations and other major locations that impact the soldier's safety.

The additional document management data would be used to move beyond simple treatment substantiation and HIPAA compliance to dramatically expand analytics and decision support for not only the deployment readiness community, but the other related business areas. Additional free text data would be used to give much more in-depth explanation of treatment stratagems and protocols and insight into individual cases thus aiding the search for health and treatment trends or patterns.

5.2 Operating Environment

This section provides an overview of the hardware and software architecture within which the Big Data applications/tools will execute in order to support their intended usage. It should provide critical insight into whether MODS can support Big Data, and if so, how the Big Data will be stored and accessed. This will frequently give meaningful information as to the basic Big Data approaches being employed by the initiative.

The MODS data warehouse is being built using Microsoft's SQL Server Parallel Data Warehouse (PDW) Massively Parallel Processing (MPP) architecture. The architecture for such a warehouse would look something like that documented in Garrett Edmondson's Blog Post [16] (see Figure 2).

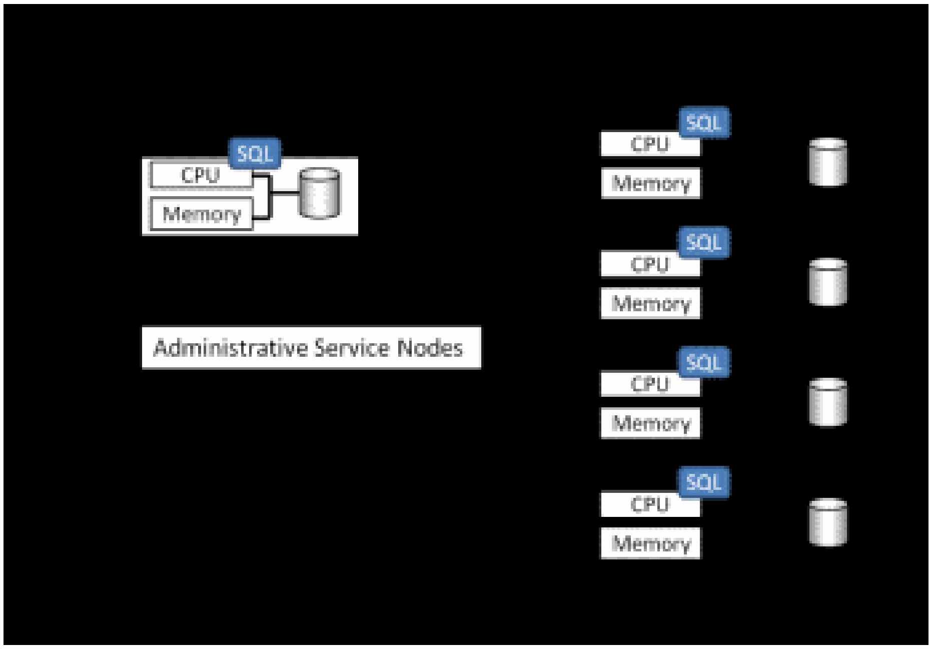


Figure 2. MS SQL Server PDW Architecture

As he states:

> Microsoft's SQL Server Parallel Data Warehouse distributes data and processing across many servers, or nodes, each of which has its own memory and disk so they can share the workload.
>
> This approach, known as massively parallel processing (MPP), has been around for several decades and is the basis for many of the largest super computers in existence today. Due to their historical high cost and complexity, MPP systems have historically been used by the largest companies and governmental organizations.
>
> This massively parallel architecture lies at the heart of Microsoft's Parallel Data Warehouse (PDW) system. PDW is a Microsoft SQL Server product designed to scale data warehouses from tens to hundreds of terabytes of data. It delivers the MPP architecture using an "appliance" model, providing preconfigured, optimized commodity hardware and software and a single point of support.
>
> Microsoft SQL Server PDW applies MPP to allow users to interactively visualize structured and unstructured (Hadoop) data.

The MODS architecture is expected to benefit from this cost-effective, complete data warehouse solution. MODS also expects to achieve future benefits that may include integration with Hadoop (Big Data), and with Microsoft Business Intelligence, as well as the Microsoft data preparation and data cleansing solutions described below [17].

5.3 Data Preparation

In order for the data to be usable in any given application (e.g., for querying and reporting, or for various forms of data mining and analytics), it must first be captured, collected, converted, edited, augmented, indexed, and loaded into the application. Data preparation is needed for both first time set up of the application to initially load the application with the data it will operate on, as well as ongoing operation of the application during which data is updated and moved into and out of the application and its operating environment. In many applications, this is called the ETL process.

MODS intends to employ Microsoft SQL Server Integration Services (SSIS) [18] to provide its ETL services. Microsoft SSIS is a platform for building enterprise-level data integration and data transformation solutions. MODS can use SSIS to solve complex business problems by copying or downloading files, sending email messages in response to events, updating data warehouses, cleansing and mining data, and managing SQL Server objects and data. SSIS can work alone or in concert with other packages to address complex business needs. SSIS can extract and transform data from a wide variety of sources such as XML data files, flat files, and relational data sources, and then load the data into one or more destinations.

SSIS includes a set of built-in tasks and transformations; tools for constructing packages to perform these tasks and transformations; and a service for running and managing these packages. MODS can use the graphical components of SSIS to create solutions without writing a single line of code; or MODS can program against the SSIS object model to create packages programmatically, or just code custom tasks and other package objects.

5.4 Data Governance

The Data Management Association (DAMA) Guide to the Data Management Body of Knowledge (DMBOK) defines data governance as “the exercise of authority and control (planning, monitoring and enforcement) over the management of data assets” [8]. This section discusses what approaches have been adopted for the management and governance of the Big Data collection.

The MODS concept of data management [19] anticipates the creation of a data governance council led by business people that will oversee the organization’s data as valuable assets. Important focus areas will be data stewardship, data standards, meta data management, master data management, data quality management, as well as risk management involving privacy and security. This activity, including decisions regarding data sensitivity, trade-offs, and legal and regulatory requirements, will itself need to be supported by a full set of analytics.

While data governance for MODS has not yet been established, the planned data governance implementation described above is consistent with DAMA’s recommendations. The question remains whether potential growth into the Big Data arena will pose significant challenges to this traditionally sound data management strategy.

One immediate area with which many organizations involved in Big Data have had to grapple is the trade-off between the value or potential value of having much more data available for analysis and the cost of setting up and operating environments for storing, processing, and managing massive amounts of new data. For example, McKinsey [35] considers this a strategic planning problem, and recommends establishing a cross-cutting strategic dialog at senior levels involving development of a plan whereby investment priorities that balance cost, speed and acceptance are sorted out through descriptions of the data to be involved, the analytic models to be employed, and the tools to be used to ensure user engagement. A similar three step approach is advocated by AT Kearney focusing on identifying big data impact areas, laying out cost scenarios, and then identifying the benefits [36].

6 Data Quality

This section discusses the quality issues with the data. Quality is frequently defined as: "Fit for purpose." Correspondingly, good quality data can be defined as: "Data that is fit for its use." Thus, the quality of the data is a function of both the data and how the data will be used.

6.1 Data Quality Dimensions

Good quality data exhibits a number of characteristics: accurate, precise, complete, consistent, timely, and authoritative. Each of these characteristics constitutes a dimension by which the quality of the data may be measured [20]. Typical definitions of these dimensions as they apply to traditional-sized data sets are as follows:

- Accuracy:
  - Degree to which data correctly reflects the real world or a surrogate source (Correctness)
  - Degree to which reported information values conform with the true or accepted values (Fitness)
- Consistency/Validity:
  - Degree to which data is synchronized across all sources
  - Degree of freedom from variation or contradiction (historical/referential/corroborative)
  - Degree of satisfaction of constraints (including syntax/format/semantics)
- Completeness/Brevity:
  - Degree to which expected records are present, and data attributes are populated
  - Degree to which duplicate entities are identified and appropriately resolved
  - Degree to which values not needed for decision making are excluded
- Timeliness:
  - Time/Utility; Degree to which currentness of data values renders them useful
  - Degree to which specified data values are up to date between data change and processing
- Pedigree/Lineage/Provenance :
  - History of data origin and subsequent ownership and transformation for traceability
- Precision/Certainty:
  - Level of detail or exactness (significant digits, rounding, truncation, resolution, sampling rate, etc.)
  - Confidence in value (vs. imprecise, approximate, uncertain, probabilistic, or fuzzy)

For a given Big Data collection, all, some, or none of these traditional dimensions may apply, or the dimensions may manifest themselves differently, or there may be a slightly different set of dimensions that the community has chosen to focus on. The intent of this case study is to determine the specifics of this situation as it might relate to MODS.

MODS decided to utilize the Federal Data Architecture Subcommittee (DAS) Data Quality

Framework for its perspective on data quality [19]. The DAS Framework defines data quality as:

> "The state of excellence that exists when data is relevant to its intended uses, and is of sufficient detail and quantity, with a high degree of accuracy and completeness, consistent with other sources, and presented in appropriate ways" [21].

Known data quality issues and considerations associated with the structured data to be loaded into MODS fall into a number of different dimensions which they call **data quality related factors**: [22]

- Data Accuracy
- Data Precision/Certainty
- Data Consistency/Validity
- Data Completeness/Brevity
- Data Non-Redundancy
- Data Availability
- Data Timeliness
- Data Pedigree/Lineage/Provenance

This list was substantially drawn from the data quality dimensions introduced at the beginning of this section with the addition of Data Non-Redundancy and Data Availability.

- Data Non-Redundancy: Data is captured in one place only, with no duplication in other systems or databases, except where redundancy is designed and controlled proactively (e.g., for performance or availability reasons).
- Data Availability: Data that is needed is captured and made available to users.

In addition, data timeliness was further described in terms of specific data characteristic requirements or needs as follows:

The following [data timeliness] needs were discussed and defined for the current MODS Data Warehouse (DW) environment:

- Base-level need is to get an accurate view of Medical Readiness and profile data to an end-of-month level.
- However, the primary value will come from having more frequent updates (and therefore more timely data) at a weekly level. For example, weekly data would be vital to support a Pandemic.
- Eventually, we need weekly updates, monthly reports, and quarterly discussion s at command level.

A set of known data quality issues were then mapped to the described data quality related factors listed above. These are presented in Appendix D.

6.2 Special Big Data Quality Issues and Considerations

There are special circumstances or conditions that must be addressed before data can be considered usable, especially related to any potential Big Data sets to be introduced for MODS.

6.2.1 Geospatial Data Quality

In its early history, geospatial data was generally of low-accuracy, and production involved many months and involved high costs. But today, technology is enabling the production of high-accuracy geospatial data at a much cheaper cost. The expansion and growth of the GIS industry in recent times is more a direct result of accurate and timely geospatial data. MODS could take advantage of this evolution [13].

The GIS data currently available for MODS is believed to be of poor quality, although it has not actually been measured. The data has low accuracy because it has been entered manually by data entry clerks. Format consistency also varies greatly due to the large number of ways an individual location can be represented.

Perhaps the biggest concern is with precision matched to the specific privacy and security requirements of the user [19]. Geospatial data reveals a lot of excellent information. If not properly controlled, access to this information could be greatly abused or misused by the country's enemies. Due to the sensitive nature of the data within the MODS system, the current MODS environment is required to comply with federal laws governing privacy and security requirements for the highly sensitive Personally Identifiable Information (PII), Protected Health Information (PHI) and Individually Identifiable Health Information (IIHI) contained within the system.[2]

During discussions with current MODS contractors, there was an expressed potential interest in utilization of devices to capture geospatial information in a more automated fashion. While there is currently little understanding of individual device reliability and precision, given industry experience, automated device capture of geospatial data would probably dramatically improve the quality of the data along the accuracy, consistency, and precision dimensions.

6.2.2 Document & Records Data Quality

For the most part, document and records data is very dependent on the meta data that is associated with each item under management. This meta data provides the identifiers, locations, status, dates, and times associated with the items as needed to permit proper filing, retrieval, and general management. For MODS, this meta data is reasonably reliable. Unreliable meta data is rare because mis-entered information is quickly captured during the many steps of the long and involved medical care and treatment process, and the extreme pressures generated by the deployment cycles levied on today's military personnel (see Section 4.2, Variety). This is not to be confused with the quality of the actual content of the documents and records which can and does vary dramatically.

The process that is used to manage documents and records meta data is called **data curation**. As defined in Wikipedia, "data curation is a term used to indicate management activities required to maintain research data long-term such that it is available for reuse and preservation." [23] As defined by Stonebraker, et.al., "[d]ata curation is the act of discovering a data source(s) of interest, cleaning and transforming the new [target] data, semantically integrating it with other local data sources, and deduplicating the resulting composite." [2] Generally, data curation as currently implemented suffers from two major problems: 1) lack of integration of all of the differing activities of the curation process end-to-end, and 2) inability of current manual curation

[2] Note that PHI and IIHI are subject to the same legal and regulatory requirements as PII in addition to specific health information laws and regulations. Due to the additional requirements applicable to PHI/IIHI, for purposes of this document PII and PHI/IIHI are treated as separate classifications of data.

approaches to scale to handle really large volumes of data.

MODS has devolved a relatively effective mechanism to curate their current collection of document and records data. However, this approach is tuned to the current scale of operations and is very manual. As MODS scales out to handle much more document and records data, it is expected that there will be unsustainable pressure placed on the current manual-based mechanisms used to cope with quality issues. It is doubtful whether these processes will be able to handle the projected new volume and velocity demands. MODS must be wary of this scaling problem.

Another problem encountered with document and records data are the legal and compliance requirements that accompany their proper management. These involve issues such as: "[r]etaining the right assets, disposing of assets accurately when their retention period is up, and being able to convince a court or regulatory body that you have this process under control. These are ... not optional elements that can be fashionably discarded." [24] In the view of some Big Data advocates, document and records data is simply a collection of tiny assets with limited individual and intrinsic value. This may be true when conducting large scale trend analyses and pattern discovery activities, but doesn't help people trying to find an email or word document that gets them out of a lawsuit or a regulatory non-compliance finding.

For MODS, this situation is most readily embodied in their HIPAA compliance requirements.

6.2.3 Text Data Quality

Text data can be thought of as information "encoded" in natural language. Like any other source of data, there are mistakes – possibly more mistakes simply because there are more words. And text is often used because it is a more efficient way to communicate information (between humans!) than through the use of detailed controlled vocabularies or other representations that would be required to encode this complex nuanced information (like a patient's symptoms or medical history).

Textual data derived from written or transcribed sources is inherently subject to data quality problems. As Brekke has stated, "[t]ext, as opposed to most other enterprise data, is very dirty data." [25] According to Seth Grimes [26]:

> "... orthodox quality thinking doesn't apply to text. It is (currently) impossible, with text, to achieve anything near the 100% definitional precision demanded by data quality purists. The problem is not only that quality steps designed for data in and from transactional and operational systems don't extend to text sources. (I'm referring to data profiling, cleansing, and standardization with a central role for master data management and data governance.) Documents are different from databases to the point where conventional data quality steps may even be undesirable in work[ing] with text. Unlike data neatly stored in database fields, text-sourced data is ambiguous. Meaning is contextually dependent and often, further, is best construed in light of user intent. There are few absolutes.

MODS textual data will be subject to the types of quality issues described above. First there are simple manual data entry errors. People make mistakes, and people's natural language mistakes are difficult to detect and correct. The program would have to go to great lengths to detect and correct such errors – and any automated cleansing might well introduce new errors. .Then there are semantic based errors; i.e., potential errors from misinterpreting the meaning in the doctor, nurse, clinical, or case management notes that have been provided. The note text is tailored to its intended primary use, which makes its interpretation heavily dependent on this context; taken out of context, the meaning may be ambiguous and underspecified. It might be possible to determine meaning probabilistically with some level of confidence using various statistical tools. It may be possible to increase this confidence by combining information from multiple sources. However, the information extracted from text will always need to be treated differently than structured data.

Most data cleansing methods are not appropriate for textual data. For Big Data collections of text, even limited data cleansing becomes impractical. Even if detected by the data warehouse, it may be impractical for the data warehouse to return the dirty textual data to the authoritative source and wait for it to be corrected before the warehouse can proceed with its business intelligence support functions. In many cases, the best that might be expected is to accept whatever quality levels are present, and just deal with the level of uncertainty involved. It will be important to really understand the use cases for this material to properly establish how it should be managed. It is critical to define what the use cases are – what a user needs to get out of the system "at scale." The Agichtein article [29] is addressing very different use cases than what might be appropriate here – where you may want population information about specific medical events; or you may need to trace what has happened to an individual over time. This is very different from mining a repository for whatever nuggets you can glean from the web at large (the Agichtein scenario). If specific passages or manageable subsets of importance can be isolated for which high quality is critical, then waiting for the authoritative sources to undertake limited cleansing efforts using traditional techniques might be reasonable. Larger scale efforts may require a more innovative DQ management strategy, for example, using an evidence-based probabilistic model that can combine multiple sources of (noisy) information; or a larger cleansing (and interpretation) effort such as crowd sourcing (for non-sensitive data). Given these realities, a decision needs to be made about whether the value of information extracted from text is less than or greater than the cost to extract and deal with data quality issues. This will dictate whether and what approaches can be pursued.

7 Data Quality Management

This section discusses how data quality issues are discovered and handled. Data management is a critical component to an organization's overall success. AT Kearney references an earlier blog which states that "poor data management can cost up to 35% of a business's operating revenue [36].

The DMBOK defines Data Quality Management as: "[p]lanning, implementation and control activities that apply quality management techniques to measure, assess, improve and assure the fitness of data for use." [8] Figure 3 presents a framework for understanding how data quality management should be viewed within a systems context [20].

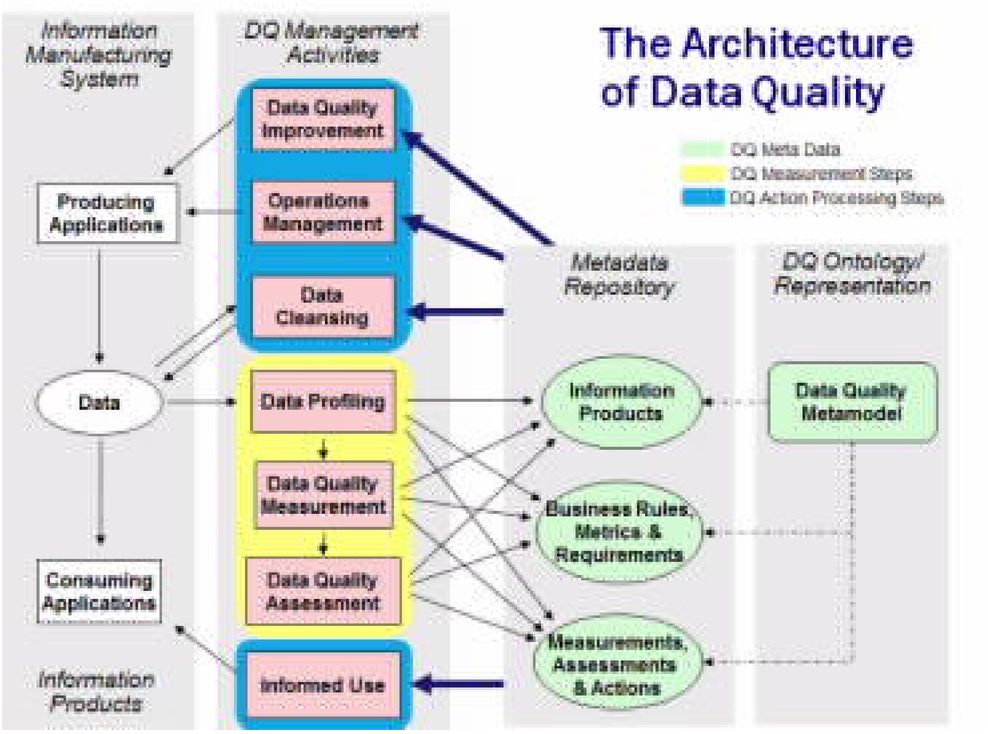


Figure 3. The Architecture of Data Quality

This figure provides a generic representation of an information manufacturing system characterized as a set of applications that produce data in the form of information products that are consumed by other applications. DQ management within this construct is accomplished through execution of two categories of activities: DQ measurement and DQ action processing.

DQ Measurement – The first category consists of three basic tasks: profiling, measurement, assessment.

> **Data Profiling** – The first step in managing the quality of the data is to profile it. Data is profiled by running it against a set of business rules to identify business rule violations. These business rule violations then become the basic building blocks from which to determine and act on DQ levels.
>
> **Data Quality Measurement** – Next, accumulations of specific business rule violations should be used as inputs to predefined DQ metrics. These metrics are the formulas set up for individual DQ dimensions such as accuracy, completeness, timeliness, etc.

Calculation of these metric formulas for specific dimensions using actual business rule violations will provide a DQ measurement for that DQ dimension.

Data Quality Assessment – Then, the DQ levels measured in the data should be compared against those DQ levels (or thresholds) expected by the user. Thresholds would be set for each individual usage context (user community or consuming system requirements).

DQ Action Processing – Finally, it is reasonable to take action based on the assessed levels of quality. There are four basic types of actions that can be taken:

Informed Use – Make the consumer aware of the quality of the data he/she are about to consume so that he/she can factor that into their decision making or analytical processes. This might entail publishing a report or updating a dashboard. It might also involve sending an email notification, raising an event flag for capture and processing by a workflow engine , or some other mechanism.

Data Cleansing – Correct the data in place so that it can be consumed by the consuming system . Sometimes, the urgency is so great that this is preferred. But it should be a last resort.

Data Operations Management – Many DQ problems are an indication of a broken information manufacturing system or operational deficiency that can be rapidly fixed , and the data regenerated.

DQ Improvement – Occasionally, the organization may choose to e ngage in a directed six-sigma, continuous improvement activity to address repeatable DQ issues so that they will be prevented from occurring in the first place .

In addition to the activities and meta data involved in data quality processing described above, data quality management also involves a number of other topics including data quality awareness, policies, organizational constructs, roles, responsibilities, training, requirements, business rules, etc., that relate to the management of quality in the various phases in the lifecycle of the data.

7.1 MODS Data Quality Management

As stated previously, MODS has decided to utilize the Federal DAS Data Quality Framework for guidance on data quality management. From the MODS concepts of data management: [19] "the data user is most concerned about data quality. However, data quality is everyone's responsibility, as it migrates from data provider to data user. It is like a chain that can be broken at any point."

MODS further describes their perspective on data quality management for the MODS structured data as follows: [19]

> The MODS DW occupies a last step in the data migration. Still, the MODS DW is responsible for insuring that the data provided to its users is of the highest quality. The mechanism to detect errors (i.e., unacceptable data as per the MODS DW requirements) is initiated in the ETL process. Due to the variety of data sources, data consistency, and data standardization are some of the factors. Enforcing standardization is one responsibility of the data governance council.
>
> In addition, data analysts will use data quality and BI tools to identify data quality

problems. Data stewards will be empowered to set data quality goals, processes, and metrics for data quality improvement initiatives. Data quality tools include the capability to profile data by capturing statistics that highlight data quality issues and provide insight into the quality of their data.

Processes for data quality include:

- Establishing processes for data analysis and feedback to data sources
- Establishing data quality benchmarks
- Establishing data quality matrices
- Establishing a continuous compliance monitoring

As with data governance for MODS, data quality management has not yet been implemented. However, if MODS adheres to the intentions stated above, the basics for good data quality management will have been established and should be operationally successful. This is entirely consistent with the Architecture of Data Quality Framework introduced above.

Furthermore, since MODS will be employing Kimball Data Warehousing methodologies, it is assumed they will be employing some of the data stewardship practices promoted in that methodology [30]. In particular, the data stewards do not change the data, but are expected to:

> Comply with corporate and regulatory policies to verify data quality, accuracy and reliability, including establishing validation procedures to be performed after each data load and prior to its release to the business. Stewards must withhold new data and communicate status if significant errors are identified.

The question is whether this basis for data quality management of traditional structured datasets will prove adequate for handling expected data quality issues encountered in the kinds of Big Data that have been highlighted as possibilities for MODS. Just managing Big Data and Big Data Quality issues in the same way as traditional sized data sets does not seem to be reasonable.

First, the amount of effort and attention required for traditional data quality management is not insignificant. If the same level of data quality is intended to be maintained in the Big Data sets, then the data quality management effort must be increased correspondingly. This by itself may be unsupportable.

Geospatial Data - While the generation and handling of large quantities of geospatial data will itself be challenging, the quality of this data should be an improvement over what is in place now. However, management of higher levels of precision while establishing and managing matching access privileges will be challenging. This is more of a traditional privacy and security management challenge related to large scale identification and authorization. Information Security Management and Privacy Management are frequently considered the most or among the most important issues affecting IT strategy, investment and implementation [37]. As called out in Finding 3 and discussed elsewhere in the document, because of the potential huge increase in the scale of the data to be protected (there is much more to secure, and the consequences of a breach are much greater), the scalability of off-the-shelf solutions to address the issues are hard to come by. While this report does not directly address this topic, it remains a fundamental area of concern for Big Data initiatives.

Document and Records Data - For document and records data, at a minimum, MODS will need to refocus its attention on meta data management and associated improvements in the capture and tracking of lineage/pedigree/provenance information. As called out in Findings 4, 5 and 6, current regulatory requirements will not go away, and existing management techniques will demand a level of curation technique that is only just now evolving in academia and industry.

Textual Data - Expansion into human language technology and NLP is not an easy step to take. The amount of energy that would need to be expended to manage the accuracy and consistency issues that will be encountered would be very high. It will be critical to understand the use cases – textual data is important if it is the main information source for something. Then there are graded techniques that can be applied to access it, ranging from indexing of keywords (and possibly concepts) – to return information to a human – to automated extraction, which has quantifiable error rates. The reason you would use textual data is precisely that it isn't available elsewhere, or that it complements information available from structured data. This is particularly true for patient records, where it is well known that retrieval using only coded information is highly unreliable.

The potential value to be gained from access to the information locked in textual data stores is expected to be very high in particular scenarios. This value must be able to justify the expenditure to obtain it and manage it, as well as, deal with any false positives and negatives that could be included in the retrieved results.

7.2 Big Data Quality Tools

MODS intends to employ the Microsoft SSIS capability they will be using for data preparation to provide data quality management functionality. One of the components of SSIS is the Data Quality Service (DQS). SSIS DQS will be used to profile, measure, and assess the quality of the various data sets feeding into and being generated by MODS. SSIS DQS will also be used to manage any data quality actions that are determined to be necessary. The SSIS DQS is described in Appendix C.

For MODS, SSIS DQS is intended to work closely with Microsoft's SQL Server PDW MPP architecture. It remains to be seen whether DQS will be adequate to handle the demands of any Big Data expansions that MODS sees as possible in the areas of Geospatial, Document & Records, and Text data.

8 References

1. J. Butler, "Big Data and Information Quality in the IRS Research Community," *2012 MIT CDO and Information Quality Symposium*, Massachusetts Institute of Technology, Cambridge, MA, July 17-19, 2012.

2. M. Stonebraker, et.al. "Data Curation at Scale: The Data Tamer System," *CIDR 2013, 6th Biennial Conference on Innovative Data Systems Research (CIDR '13)*, Asilomar, California, January 6-9, 2013. Available at: http://www.cidrdb.org/cidr2013/Papers/CIDR13_Paper28.pdf

3. R. Johnson and R. Zahavi, "Traditional Data Warehousing Meets Big Data: What Does It Mean for the Enterprise?" *CUTTER IT JOURNAL*, *Cutter Information LLC*, October 2012, ©2012.

4. "Technical Specification ISO/TS 8000-110:2008(E), Data Quality – Part 110," "Master Data: Exchange of Characteristic Data: Syntax, Semantic Encoding, and Conformance to Data Specification," First edition, 2008, *International Organization for Standardization (ISO)*, Switzerland.

5. "Analyze & Decide, Enabling The End User," MS PowerPoint Briefing, © 2012 Oracle Corporation, October 2012.

6. W. H. Inmon, A. Nesavich, "Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence," Prentice-Hall, Pearson Education, Boston, MA, 2008.

7. J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, "Big Data for Dummies," John Wiley & Sons, 2013.

8. "DAMA Guide to the Data Management Body of Knowledge," *DAMA International*, Technics Publications, Bradley Beach, NJ, 2010.

9. "Executive Order (EO) 12906," *National Spatial Data Infrastructure (NSDI)*, April 1994.

10. J. Erlichman, Editor, "Geospatial Applications in Government," *On The FrontLines*, Volume 4, Number 2, February 2012.

11. M. Brackett, "What Are Unstructured Data?" Dataversity, © 2012, *Wilshire Conferences*, 2011. Available at: http://www.dataversity.net/what-are-unstructured-data/

12. Army Medicine, "GIS Branch Helps Map Soldier Health," *Mercury July 2012, Volume 39, No. 10.* Available at: http://www.armymedicine.army.mil/news/mercury/12-07/Mercury-July2012.pdf,

13. C. Killpack, "BIG DATA, Bigger Opportunity," *Geospatial World*, April 2011.

14. A. O'Brien, "The Impact of Big Data on Government," *IDC Government Insights*, White Paper, Sponsored by Iron Mountain, #GI237315, October 2012.

15. C. Thompson, "Text Analytics – Ride the Wave or Stay Ashore?" Deloitte Consulting LLP, © 2013, Deloitte Development LLC. Available at: http://www.deloitte.com/view/en_US/us/Services/additional-services/deloitte-analytics-service/short-takes/f10873be01a1c310VgnVCM2000003356f70aRCRD.htm.

16. G. Edmondson, "Massively Parallel Processing and the Parallel Data Warehouse," Blog

Post on Microsoft Business Intelligence and Data Warehousing, April 15, 2012, Simple Talk Publishing.

17. "Appliance: Parallel Data Warehouse (PDW)," Microsoft SQL Server, ©2013. Available at: http://www.microsoft.com/en-us/sqlserver/solutions-technologies/data-warehousing/pdw.aspx

18. "SQL Server Integration Services," Microsoft Developer Network, ©2013. Available at: http://msdn.microsoft.com/en-us/library/ms141026.aspx

19. "U.S. Army Medical Command (MEDCOM) Medical Operational Data System Data Warehouse Environment (MODS DW), Concept of Operations (CONOPs)," Version 1.0, June 25, 2011.

20. D. K. Becker, "An Extended Data Model for Data Quality," MS PowerPoint Briefing, The MITRE Corporation, Bedford, MA, 2012.

21. "Applying Proven Quality Principles to Data Sources," Federal DAS Data Quality Framework, Version 1.0, October 1, 2008.

22. "Business Intelligence (BI) Requirements, MEDCOM MODS Data Warehouse, Version 2.1," Prepared by the MITRE Corporation for U.S. Army Office of the Surgeon General Medical Readiness, April 23, 2013.

23. "Data Curation" *Wikipedia, The Free Encyclopedia*. Wikimedia Foundation, Inc., as modified on February 15, 2013 at 14:21, and retrieved on May 2, 2013.

24. M. Alsup, "Is Traditional Records Management Dead?" Gimmal Group, © AIIM 2013, July 01, 2012. Available at: http://www.aiim.org/community/blogs/expert/Is-Records-Management-Dead

25. V. Brekke, "Understanding and Evaluating Big Data Text Analytics Solutions," *Social Intent,* October 24, 2012. Available at: http://blog.socialintent.com/2012/11/understanding-and-evaluating-text-analytics-solutions-for-big-data/

26. S. Grimes, "Text Data Quality," TechTarget, BeyeNETWORK, November 17, 2009. Available at: http://www.b-eye-network.com/view/12072

27. B. Knight, D. Knight, M. Davis and W.Snyder, "Knight's Microsoft SQL Server 2012 Integration Services 24-Hour Trainer," Wrox Press, 2013.

28. "Microsoft Big Data Solution Sheet," ©2011 Microsoft Corporation. Available at: www.microsoft.com/bigdata

29. E. Agichtein, "Scaling Information Extraction to Large Document Collections," Microsoft Research, Bulletin of the IEEE Computer Society, 2005. Available at: http://www.mathcs.emory.edu/~eugene/papers/DEB05-agichtein.pdf

30. B. Becker, "Data Stewardship 101: First Step to Quality and Consistency," Business Intelligence and Data Warehouse Articles, The Kimball Group. Available at: http://www.kimballgroup.com/2006/06/01/data-stewardship-101-first-step-to-quality-and-consistency/.

31. R. DeFrangesco, "Comparing Comprehensive Solutions vs. Best of Breed," IT Inside Out, June 30, 2009. Available at: http://www.itbusinessedge.com/cm/blogs/defrangesco/comparing-comprehensive-

solutions-vs-best-of-breed/?cs=33735, retrieved on August 3, 2013, Copyright © 2012 Quin Street, Inc. All rights reserved.

32. B. Deeter,“Best of Breed Applications Finally Have Their Day,” CIO Journal, February 14, 2013. Available at: http://blogs.wsj.com/cio/2013/02/14/best-of-breed-applications-finally-have-their-day/. Retrieved on August 3, 2013, Copyright ©2013 Dow Jones & Company, Inc. All Rights Reserved.

33. T. Huseby, M. Sansone, K. Heung, “Is It the Beginning of the End for Best of Breed,” ATKearney, 2012. Available at: https://www.atkearney.com/strategic-it/ideas-insights/article/-/asset_publisher/LCcgOeS4t85g/content/is-it-the-beginning-of-the-end-for-best-of-breed-/10192. Retrieved on August 3, 2013, A.T. Kearney, Inc. All Rights Reserved.

34. K. Cukier, V. Mayer-Schoenberger, “ The Rise of Big Data,” Council on Foreign Affairs, May/June 2013. Available at: http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data . Retrieved on May 6, 2013, Copyright © 2002-2013 by the Council on Foreign Relations, Inc. All rights reserved.

35. S. Biesdorf, D. Court, and P. Willmott, “Big Data: What’s Your Plan?”, McKinsey Quarterly, March 2013. Available at: http://www.mckinsey.com/insights/business_technology/big_data_whats_your_plan, Retrieved on August 5, 2012, © 1996-2013 McKinsey & Company.

36. C. Hagen, et.al., “Big Data and the Creative Destruction of Today’s Business Models,” ATKearney, 2012. Available at: https://www.atkearney.com/documents/10192/698536/Big+Data+and+the+Creative+Destruction+of+Todays+Business+Models.pdf/f05aed38-6c26-431d-8500-d75a2c384919 . Retrieved on August 5, 2013, A.T. Kearney, Inc. All Rights Reserved.

37. “’Information Security Management’ Is Top Tech Issue for 7th Year in a Row,” Public CIO, January 13, 2009. Available at: http://www.govtech.com/pcio/Information-Security-Management-is-Top-Tech.html. Retrieved on August 5, 2013, Government Technology.

Appendix A Acronyms

AHLTA	Armed Forces Health Longitudinal Technology Application
AMEDD	Army Medical Department
BI	Business Intelligence
BIOs	Business Improvement Opportunities
BQs	Business Questions
CONOPS	Concept of Operations
CRUD	Create, Read, Update and Delete
DAMA	Data Management Association
DAS	Data Architecture Subcommittee
DHA	Deployment Health Assessment
DMBOK	Data Management Body of Knowledge
DMHRSi	Defense Medical Human Resource System Internet
DoD	Department of Defense
DQ	Data Quality
DQKB	Data Quality Knowledge Base
DQS	Data Quality Service
DW	Data Warehouse
EDW	Enterprise Data Warehouse
ELT	Extract, Load & Transform
ETL	Extract, Transform & Load
GCSS-AF	Global Combat Support System – Air Force
GIS	Geographic Information System
GPS	Global Positioning Satellite
HIPAA	Health Insurance Portability and Accountability Act
HR	Human Resources
HRC	Human Resource Command
IIHI	Individually Identifiable Health Information
LOE	Levels of Effort
MDM	Master Data Management
MEDCOM	U. S. Army Medical Command
MHS	Military Health System
MNR	Medical Not Ready
MODS	Medical Operational Data System
MOIE	Mission Oriented Investigation and Experimentation
MPP	Massively Parallel Processing
NLP	Natural Language Processing
OCR	Optical Character Recognition
ODBC	Open Database Connectivity
OLAP	Online Analytical Processing
OLTP	Online Transaction Processing
PDW	Parallel Data Warehouse
PHI	Protected Health Information
PII	Personally Identifiable Information
SME	Subject Matter Expert
SRP	Soldier Readiness Processing
SSIS	Server Integration Services
TAPDB	Total Army Personnel Database

UIC	Unit Identification Code
WT	Warrior Transition

Appendix B **Technology Description**

B.1 Data Warehousing

A data warehouse is a centralized collection of current and historical data typically used for reporting and analysis. It is aggregated from multiple operational sources into a centralized repository where it is normalized for common query and access. It frequently utilizes specialized hardware and software necessary to optimize the movement of data into and out of the repository, as well as, to support search queries across the very large collection of data.

The GCSS-AF Data Services Enterprise Data Warehouse (EDW) implements a generic data warehouse architecture that is depicted in Figure 4. The figure specifies a number of specific components that constitute EDW.

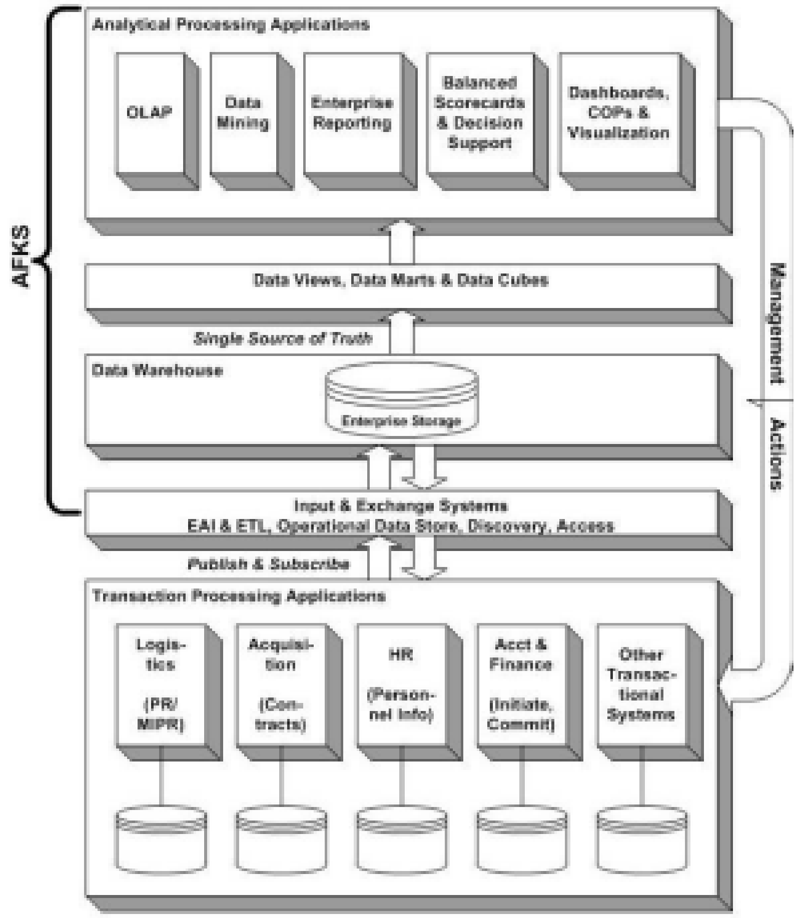

Analytical Processing Applications
OLAP
Data Mining
Enterprise Reporting
Balanced Scorecards & Decision Support
Dashboards, COPs & Visualization
AFKS
Data Views, Data Marts & Data Cubes
Single Source of Truth
Data Warehouse
Enterprise Storage
Management
Actions
Input & Exchange Systems
EAI & ETL, Operational Data Store, Discovery, Access
Publish & Subscribe
Transaction Processing Applications
Logis-tics (PR/MIPR)
Acquisi-tion (Con-tracts)
HR (Person-nel Info)
Acct & Finance (Initiate, Commit)
Other Transac-tional Systems

Figure 4. Generic Enterprise Data Warehouse Architecture

There are currently five major components to the generic data warehouse architecture:

1. The analytical processing applications that are the primary use of the data warehouse (often called business intelligence or BI).
2. The derivatives of the data warehouse that are used to support the analytical processing applications.

3. The data warehouse itself.

4. The input and exchange systems that feed data to and extract data from the data warehouse.

5. The transactional processing applications that are the authoritative sources of the data in the data warehouse.

On the input side, the primary flow is from the transactional processing applications, through the input systems to the data warehouse. On the output side, when the data warehouse is being used as the single source of truth for all analytical processing, the primary flow is from the data warehouse, through the derivative forms to the analytical processing applications. Also on the output side, when the data warehouse is being used as an archive or broker for transactional processing, the primary flow is from the data warehouse, through the exchange systems to the transactional processing applications.

The EDW provides implementations for four of these five major architectural components, all but the transactional processing applications. These are essentially a set of tools and capabilities that a program can pick and choose from when attempting to understand how to incorporate BI and the EDW into their architectural evolution.

B.1.1 Analytical vs. Transactional Processing Applications

The primary distinction that should be drawn from Figure 4 is the separation of all analytical processing applications from all transactional processing applications. This separation is extremely important because it is the fundamental basis for the emergence and evolution of data warehousing as a separate computing concept. It represents a basic attempt to separate two entirely different sets of users: those who are doing primarily analysis, and those who are primarily processing transactions (e.g., those who manage work, and those who do work).

Analytical users require mostly read-only access to the data. The BI applications they use are typically ad hoc and canned query tools, enterprise reporting, data mining, balanced scorecards, dashboards, decision support, and data visualization tools. These tools frequently look at a lot of data but hardly ever change it. This typically represents the managerial or controller activity of an organization.

Transactional users on the other hand require mostly CRUD (Create, Read, Update and Delete) access to the data. This access is typically isolated to single records or small collections of records, and it is typically focused in separate functional areas or stovepipes. The applications of transactional users are usually very specialized and configured to the type of work they are performing. This typically represents the actual work that an organization is chartered to perform.

If both of these sets of users were operating off of the same data store, they would severely interfere with each other's ability to do their respective jobs. Transactional users would frequently be locking the tables and changing the underlying data, thus inhibiting analytical users from being able to run their queries, view the enterprise data, and get their reports and analyses completed in a timely manner or on a stable basis. On the other hand, the large reports and queries and detailed analyses required by analytical users can consume huge amounts of computing resources and take a long time to run, thus interfering with the ability of transactional users to get real work done.

The answer was to separate these two very different types of applications. The transactional applications and their users would operate off of local operational data stores that could efficiently service their critical transactional processing needs. These local data stores would be optimized to support CRUD activity to give the transactional users the very best system responsiveness to their specific needs. There might be some small-scale reporting and analysis on the local data store, in order to provide some critical operational or tactical needs, but this would be kept to a minimum. As a transaction was completed, or on a convenient periodic basis (hourly, daily, weekly, etc.), the transactional data would be extracted from the local data store, and would be transformed and loaded up to the data warehouse.

The data warehouse would be structured to support the reporting and analytical BI applications. The analytical applications and their users would operate off of the enterprise data warehouse and its derivatives (data views, data marts and data cubes) that could efficiently service their critical analytical processing needs. The enterprise data warehouse would be optimized to support read only query activity to give the analytical users the very best system responsiveness to their specific needs. There would be occasional update activity to move periodic loads of new data into the data warehouse, but these could be scheduled for off peak periods when their impact would have little effect, and staging data base techniques could be used to ameliorate their effects even more.

It is vitally important for system architects to establish whether the applications that make up their system are transactional or analytical in nature. If an application is transactional, it should be structured with a local data store (such as Oracle, Sybase or SQL Server) that is optimized for the CRUD activity peculiar to its functional requirements. If the application is analytical in nature it should be strongly considered for implementation using EDW tools and the EDW architecture. EDW should not be considered for use as an operational or transaction data store to support the transactional processing applications directly. It is only optimized to support analytical processing applications.

B.2 SSIS Data Quality Services

SSIS Data Quality Services (DQS) provides some tools and services to help improve data. DQS is intended to help in the following areas: [27]

- Completeness - Are data values missing? If you have 25,000 customers and only 15,000 valid email addresses for them, your email address field is 60 percent complete.
- Consistency - Are data values being used consistently? If Gen Mgr. and GM are alternate terms that refer to the General Manager position, are position field values used consistently? (The answer is no.) Even though you know that the values refer to the same position, you must make the values consistent. This is important because you will use this data for comparison and aggregation. Use of inconsistent values provides inaccurate results.
- Conformity - If special formatting is required for certain fields, do the data values match the correct formatting? You can import data from several sources that store values with the same meaning in different ways. Consider the “gender” field. One source provides values of M and F. Another system uses 1 and 2. A third system uses Male and Female. Even if the values are consistent within each system, when brought together, there is a problem. For the “gender” field to be conformed, you must choose the correct representation, and convert the input values to the conformed, approved values.
- Validity - Do valid data values fall within acceptable ranges? For example, the definition of an “age” field should include values between 1 and 120. Negative values do not make any sense, nor do values over 120.
- Accuracy - Do the values represent the true, factual value for the object? As an example, the “age” field value of 25 for me is valid, but not accurate. I am 59.
- Duplication - Are there multiple instances of the same object? For example, two rows in the

customer table that represent the same object (customer)? Perhaps the single customer represented in two rows has two different names, like Bob Smith and Robert Smith. Maybe the customer is a woman who has changed her name due to marriage or divorce. Another common example that can yield duplicates is a customer who has moved, and you have a customer record for the old address and the new address. It is very common to end up with duplicate values for the same object when combining rows from multiple data sources; for instance, when two companies merge.

Data cleansing decisions can be made using DQS in each of the dimensions above; for example, to replace "M" and "1" with the value "Male" in the gender field. These decisions are contained in a store called the Data Quality Knowledge Base (DQKB). During cleansing of data, the information in the DQKB is used to automate parts of the process.

Data cleansing is a process, and not a destination. It is continual and iterative. It begins with manual definition and error identification. As cleansing progresses, the knowledge base grows and improves, and manual work decreases.

Use of DQS is comprised of three main steps:

- Knowledge Base Management
- Data Cleansing and Matching
- Administration and Monitoring

In the Knowledge Base Management phase, domains are created. An example of a domain is Gender. The Gender domain contains information about Gender as a class of information, and Gender should be a string data type. A list of valid values is provided that can be imported from a file or from the database. Valid values can also come from reference data in the cloud. Rules can also be created that apply to domains. The early goal is to improve the quality of the knowledge base by ensuring that the domain information is accurate and complete.

Composite domains can also be created. For example, given First Name and Last Name domains, a composite domain called Full Name can be then created that is composed of the First Name and Last Name domains. Separate, additional rules can be supplied to the composite domain.

While cleansing data in the Data Cleansing and Matching phase, incoming data is processed using the information stored in the knowledge base. Completeness and accuracy are displayed, along with any corrections and suggestions made by the knowledge base. Values can then be approved or rejected, exported as output from the process, and used as a data source for ETL loading of data.

In this way, data cleansing consists of both computer-assisted and interactive cleansing. The data is cleansed via knowledge base information, but interactively, reviewed and approved, or rejected, yielding the final output. In addition, corrected values can be provided during the interactive cleansing.

The DQS cleansing processing will automatically place data in tabs, which are described below. As your data works through interactive cleansing, changes, and corrections, data may be moved into a different tab.

- Correct - An exact match was found in the knowledge base or you approved the value.
- Corrected - Values corrected by DQS with a high confidence level, or you provided a value in the Correct to column and approved.
- Invalid - Values marked in the knowledge base as invalid , or that failed a domain rule or

reference data, or values that were rejected by you.

- New - Valid values for which there is not enough information (not marked as invalid in the knowledge base), and values for which there is a suggestion with a low confidence.
- Suggested - DQS suggests these values. The confidence level is not high enough to be Corrected, but the confidence level is above the minimum level to provide this as a suggestion. You must review and approve/reject these values. The confidence levels for Corrected and Suggested can be set by the DQS administrator.

B.3 Microsoft Big Data Solution [28]

NOTE: The following material is drawn exclusively from the "Microsoft Big Data Solution Sheet", and is incorporated here to provide an in-line description of the technology for quick reference. Some statements may not represent the views of the MITRE authors.

Microsoft's vision is to provide business insights to all users from any data, including insights previously hidden in unstructured data. To achieve this goal, Microsoft will ship an Apache Hadoop™ based distribution for Windows Server and Windows Azure to help accelerate its adoption in the Enterprise.

This new Hadoop based distribution from Microsoft enables customers to derive business insights on structured and unstructured data of any size and activate new types of data. Rich insights from Hadoop can be combined seamlessly with the Microsoft Business Intelligence Platform.

B.3.1 Key Benefits

- Broader access of Hadoop to end users, IT professionals , and developers through easy installation and configuration and simplified programming with JavaScript .
- Enterprise-ready Hadoop distribution with greater security, performance , and ease of management.
- Breakthrough insights through the use of familiar tools such as PowerPivot for Excel, SQL Server Analysis and Reporting Services .

Microsoft's Big Data solution also offers interoperability with other Hadoop distributions, enabling customers to derive insights from several sources.

- Two Hadoop Connectors: First, Microsoft offers two Hadoop connectors that enable customers to move data seamlessly between Hadoop and SQL Server or SQL Server PDW. These two Hadoop connectors are now available to existing customers.
- Hive Open Database Connectivity (ODBC) Driver, plus Excel Hive Add-In: Second, Microsoft offers a new Hive ODBC Driver and an Excel Hive Add-in that enable customers to move data from Hive directly into Excel, or Microsoft BI tools such as PowerPivot, for analysis.

B.3.2 Broadening Access to Hadoop

Microsoft is committed to broadening the accessibility and usage of Hadoop to users, developers, and IT professionals.

The new Hadoop based distribution for Windows offers IT professionals ease of use by simplifying the acquisition, installation, and configuration experience. Thanks to smart packaging of Hadoop and its toolset, customers can install and deploy Hadoop in hours instead of days. End users can use the Hive ODBC Driver or Hive Add-in for Excel to analyze data from Hadoop using familiar tools such as Microsoft Excel and award winning BI clients such as PowerPivot for Excel.

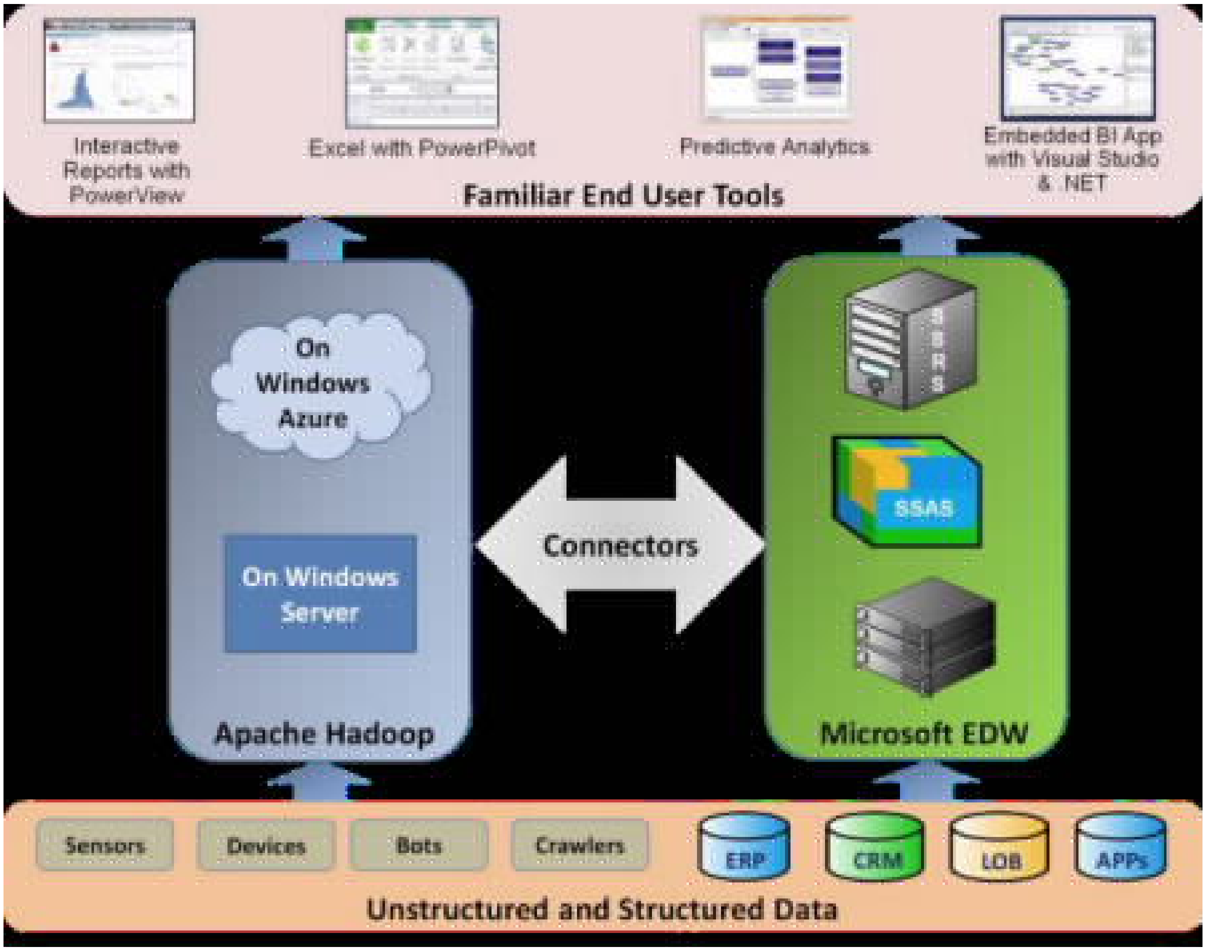


Figure 5. Outline of Microsoft's Big Data Solution

For developers, Microsoft is investing to make JavaScript a first class language within Big Data, by making it possible to write high performance Map/Reduce jobs using JavaScript. In addition, our JavaScript console will allow users to write JavaScript Map/Reduce jobs, Pig-Latin, and Hive queries from the browser to execute their Hadoop jobs. This is the sort of innovation that Microsoft hopes to contribute back as proposals to the community.

Appendix C Project Description

The following information is drawn almost entirely from two sources, MODS Concept of Operations (CONOPS) [19] and the MODS CONOPS Stakeholder Feedback Meeting [2].

C.1 Background

The Medical Operational Data System (MODS) is a Military Health Services System that provides the Army Medical Department (AMEDD) with an integrated automation system that supports all phases of Human Resource Life-Cycle Management in peacetime and during mobilization. The MODS provides commanders, staff, and functional managers of AMEDD organizations with a real-time source of information on the qualifications, training, special pay, and readiness of AMEDD personnel.

Since the early 1990s, the MODS architecture has been designed as a myriad of stove-piped two-tier, Web-based applications. Those applications were maintained and upgraded, and, in some cases, extended to expand the existing capabilities in a gradual manner. The historical MODS design has served the Army well, but the rapid growth of applications has introduced disparity, code complexity, and data redundancy, making the system inefficient and outdated by today's industry standards.

Currently, the MODS system maintains approximately 60 system interfaces, which are updated on a periodical basis. These interfaces bring in over 75 percent of the data used in the system. The MODS system moves critical assignment, training, and qualification information through those interfaces in order for decision makers and managers to work their requirements quickly. The time spent acquiring data can now be used to analyze the information.

Future design and implementations of MODS facets must be responsive to leadership and congressional ad-hoc inquiries. The system will also be required to respond to the Human Resources (HR) needs and wants of its stakeholders. Both challenges will necessitate speedy, accurate, and cost-effective data querying across its databases and applications.

C.2 System Goals and Objectives

The MODS mission is to deliver an information-management data system that provides access to reliable HR information and Army readiness response. The objective and goals of the system are:

- Enhance the analysis and decision making of Army and the Department of Defense (DoD) leaders by linking their database systems with the MODS using a standard interface.
- Support and maintain a human resource decision -management system for Army personnel accessions, promotions, training, pay, and transitions .
- Provide interfaces that feed critical , medical human-resource information to the Tri-Service Defense Medical Human Resource System internet (DMHRSi).
- Standardize the Army AMEDD systems and interfaces to interact with the other MODS applications.
- Provide tools that help automate the medical logistics management and planning.

C.3 System Description

In this section, the current architecture of the MODS is described. Figure 6 illustrates the MODS

Enterprise database design that provides the basis for data storage and retrieval of standard U.S. Army personnel data. The system receives data from the mainframe stream provided by U.S. Army Human Resources Command (HRC) where the data sources are:

- Total Army Personnel Database - Active Officer (TAPDB-AO). This is HR-related information for active-duty officer personnel only.
- Total Army Personnel Database - Active Enlisted (TAPDB-AE). This is HR-related information for active-duty enlisted personnel only.
- Total Army Personnel Database - Reserve (TAPDB-R). This is HR-related information for Army Reserve personnel.
- Total Army Personnel Database - National Guard (TAPDB-G). This is HR-related information for Army National Guard Personnel .
- Army Civilian Personnel System (ACPERS). This is HR-related information for Department of the Army (DA) Civilians.
- Data for U.S. Air Force personnel is provided by the Air Force Corporate Health Information Processing Service (AFCHIPS) personnel system.

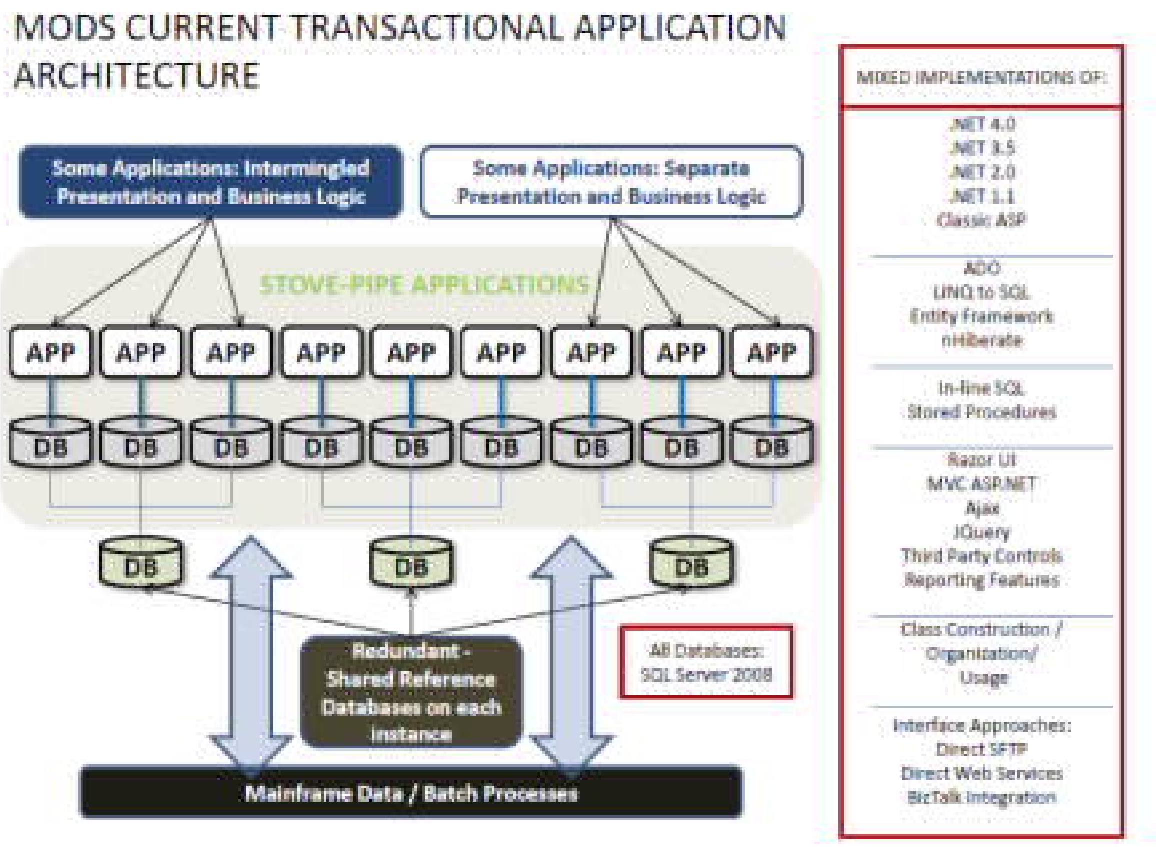

MODS CURRENT TRANSACTIONAL APPLICATION ARCHITECTURE
Some Applications: Intermingled Presentation and Business Logic
Some Applications: Separate Presentation and Business Logic
STOVE-PIPE APPLICATIONS
APP
APP
APP
APP
APP
APP
APP
APP
APP
DB
DB
DB
DB
DB
DB
DB
DB
DB
DB
DB
DB
Redundant - Shared Reference Databases on each instance
All Databases: SQL Server 2008
Mainframe Data / Batch Processes
MIXED IMPLEMENTATIONS OF:
.NET 4.0
.NET 3.5
.NET 2.0
.NET 1.1
Classic ASP
ADO
LINQ to SQL
Entity Framework
nHiberate
In-line SQL
Stored Procedures
Razor UI
MVC ASP.NET
Ajax
JQuery
Third Party Controls
Reporting Features
Class Construction / Organization/ Usage
Interface Approaches:
Direct SFTP
Direct Web Services
BizTalk Integration

Figure 6. MODS Current Transactional Application Architecture

C.4 Target Technical Environment

The MODS DW will leverage the most advanced technical platform available in data warehousing, a MPP architecture that scales to hundreds of TBs and thousands of users. At the core of Microsoft's SQL Server PDW is the ability to distribute work (e.g., data loads), user queries, and replication across many commodity hardware nodes. Each node comprises commodity Intel CPUs and data storage, all connected through a high-speed, high-bandwidth backbone. Central controllers manage the work distribution across the many nodes, delivering orders of magnitude performance improvement over non-parallel processing systems.

From a data perspective, there is no single "right" data architecture for all situations. Normative data architecture for a DW environment typically comprises a data warehouse layer, a data mart layer, and a number of secondary supporting layers, as described later in this section. However, these industry terms can create confusion as they often mean different things to different people, and can depend upon the use of certain technologies, tools, methodologies, and approaches. The two primary data layers in the MODS DW environment are depicted in Figure 7, and the related terms are defined below.

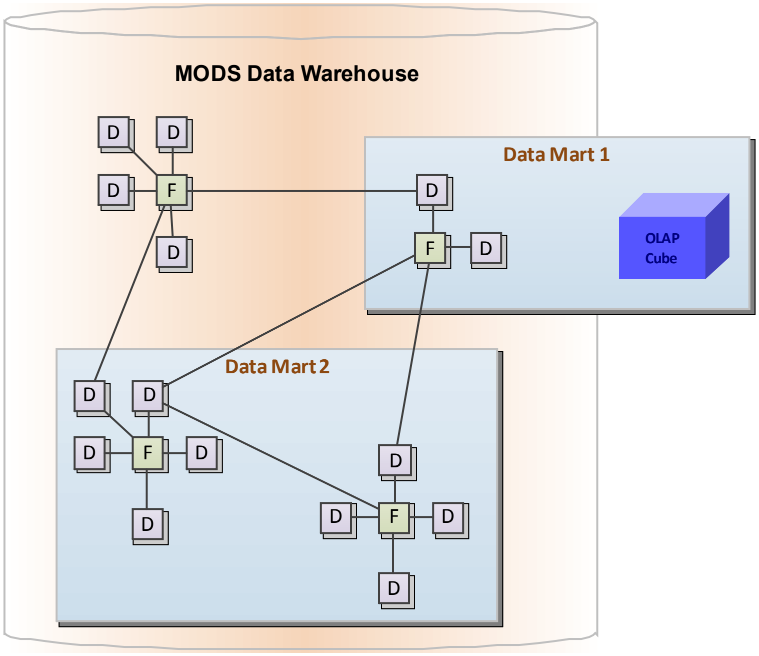


Figure 7. Depiction of MODS Data Warehouse and Data Mart Layers

- **Data Warehouse**. An integrated, physical database comprising detailed (atomic) and summary transactional data (facts), organized using dimensional schemas (as per the Kimball approach) that contains standardized (conformed) reference data (dimensions). Fact tables would capture personnel counts, assignment durations, readiness statistics, and other information derived from program activities. Conformed dimensions, such as Person, Duty Station, Occupation, and Unit will be shared across different fact tables. While the data warehouse layer is designed around the MEDCOM's key business processes, its aim is to support multiple uses and users for different purposes. This layer represents integrated, standardized, and scrubbed data across the set of the MODS application systems.
- **Data Mart**. Each data mart is a logical collection of relevant fact tables (F) and dimension tables (D) in the data warehouse layer and/or pre-defined OLAP cubes that, together, represent a holistic capability needed by a particular set of users for a specific purpose. For example, a

data mart for AMEDD HR may include multiple fact tables for analyzing the status of/need for medical certifications. Another mart comprising various fact tables might support planning future medical personnel resource needs.

C.5 Conceptual MODS Data Warehouse Architecture

Figure 8 portrays the major components of the end-to-end data warehouse environment from a conceptual perspective. Each major component (numbered 1-10) is described below.

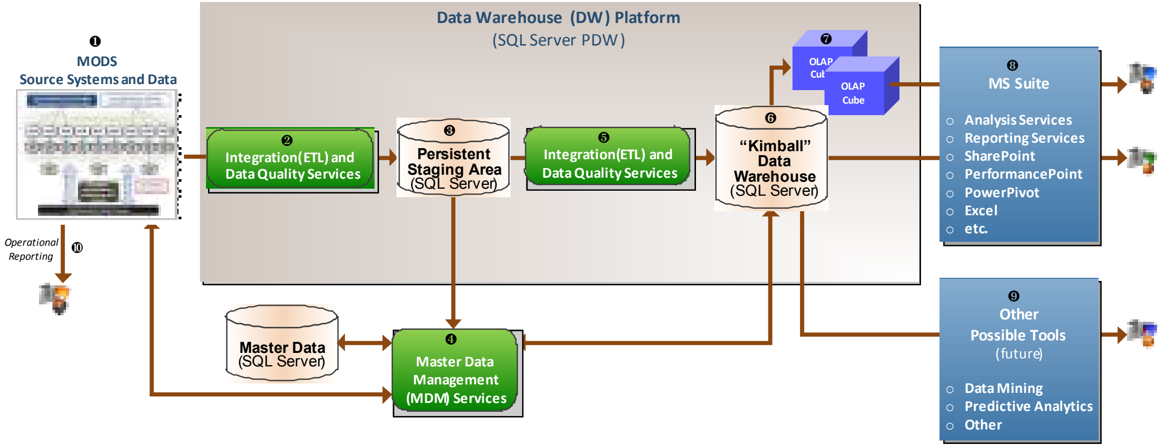


Figure 8. Conceptual MODS DW Architecture

1. **Source Data**. Pre-determined data across the MODS operational systems will be fed to the MODS data warehouse through the integration services periodically (e.g., hourly, nightly, etc.). The scope and timeliness of data feeds will be driven by business requirements.
2. **Integration** (e.g., ETL/ELT) and Data Quality Services to transform source data into Staging. These services will extract (pull) and/or receive (via a MODS source data push), and perform some data integration needed to stage the data into the warehouse environment. For example, it may be practical to create certain standardized data tables and/or data attributes from across the disparate systems in the staging area. Using ETL or ELT, processes will move the data and implement business rules for merging related source data into common structures.
3. **Persistent Staging Area**. Also, referred to as a "landing area," the staging area supports the capture of raw data, and through the ETL/ELT, processes and begins to merge and standardize data for analytical purposes. Data quality improvements include standardizing inconsistent code values, dates, and identifiers. It also includes merging, filtering, and separating physical transactions into logical records. This data layer helps to simplify the work and rules needed to integrate data into the DW layer, which will require additional transformation in component 5 below. Typically, the data structure is more "normalized" than the DW layer and provides a more direct source of data, if additional content is needed in the warehouse.
4. **Master Data Management (MDM) Services**. MDM establishes a system of record

from which other processes benefit. Initially, the vision for the MODS' MDM is to support a consistent view of person and unit data for *analytical* purposes (i.e., using the DW). Downstream use of master data will be used to improve *operational* processes within the core MODS transaction systems by re-engineering of those systems to utilize master data directly. MDM data includes:

   a. Core personnel data about Service members in the active Army, Air National Guard, Army Reserve, and some Air Force, Navy, Marine, Coast Guard data to facilitate processing of immunization data, medical readiness data, and other information;
   b. U.S. Army Unit Identification Codes and primary information about those units (i.e., location, operational status, start date, expiration date).

5. **Integration** (e.g., ETL/ELT) and Data Quality Services to transform staged data into the DW. While these services utilize the same core capabilities as those used in component 2, the use of ETL/ELT services in this component is to maintain the data warehouse. This additional functionality includes the instantiation of dimensional schemas, including conformed dimensions and slowly changing dimensions to provide historical views of attributes, such as personnel, unit, etc.
6. **MODS DW**. The MODS DW is the result of component 5 above and by applying a Kimball approach[3] to data design contains dimensional schemas as the primary organizing framework in the database. This includes highly detailed (atomic) and summarized (derived) views, including the instantiation of commonly used measures and metrics for reporting and analysis.
7. **OLAP Cubes**. OLAP represents a subset of BI capabilities that capitalizes on a further transformation of data into "cubes." These views of the DW may be logical or physical structures, depending upon architectural decisions, but, in either case, enable users to perform very fast, ad hoc, and highly dynamic cross-tabulations, such as data pivots, summarizations, graphs, charts, filters, sorts, groupings, etc.
8. **Microsoft Suite**. End users will be provided a set of tools within the Microsoft suite, depending upon their need and level of analytical sophistication required. Tools for basic "canned" reporting, ad hoc analysis, Excel-based modeling, OLAP, metrics analysis, and dashboards form the suite and leverage capabilities from Microsoft Analysis Services, Reporting Services, SharePoint, PerformancePoint, and Excel/PowerPivot.
9. **Other Possible Tools**. Looking ahead, other third-party tools, including niche capabilities for data mining and predictive analytics, can be utilized. In some cases, data from the DW may be accessible directly by those tools or may need to be extracted and placed in special segregated environments that exclusively support custom data analysis and/or modeling.
10. **Operational Reporting**. While not specifically in the scope of the DW, it is important to note that some operational reporting will be needed from within the MODS OLTP applications. As with most DW efforts, it is not likely that the analytical capabilities within the DW will support a pure operational view of the data easily. This may be caused by several factors:
    - Lags in data timeliness.
    - Data transformations needed for analytical standardization that inherently conflict with a single system's operational view.

[3] The Kimball approach to data warehousing is a commonly applied set of techniques expressed in many books, articles, and courses authored by Ralph Kimball.

- Availability of some data that resides only within the operational environment (and may have specific security provisions and protections to safeguard it from inappropriate use and/or users).

C.6 MODS Data Paradigm

The MODS data paradigm is illustrated in Figure 9. There are three components related to the paradigm. All has been fed by the soldier demographic and organizational data.

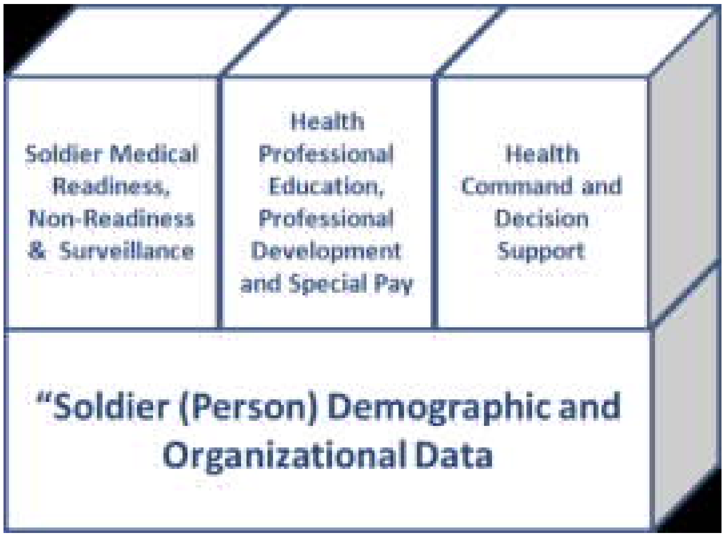

Soldier Medical Readiness, Non-Readiness & Surveillance
Health Professional Education, Professional Development and Special Pay
Health Command and Decision Support
"Soldier (Person) Demographic and Organizational Data

Figure 9. MODS Data Paradigm

Each of the components includes a set of applications:

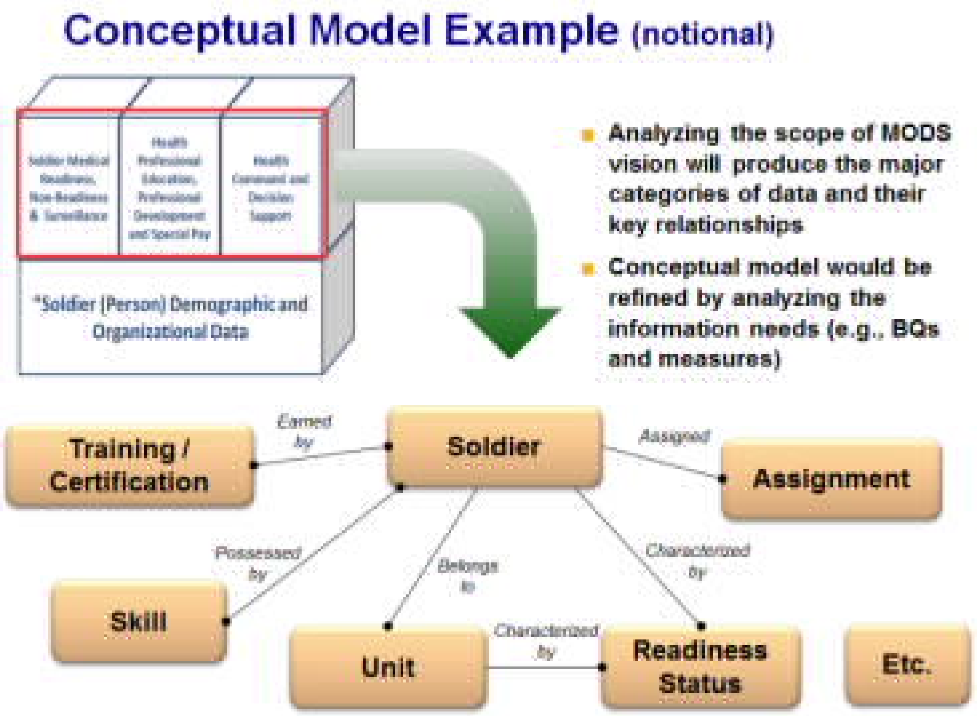
- Soldier Medical Readiness, Non-Readiness , and Surveillance:
  - Behavior Health Data Platform (BHDP)
  - Medical Protection System (MEDPROS)
  - Electronic Profile (eProfile)
  - Medical Health Assessment (MHA)
  - Personnel Blast and Containment Tracker (PBCT)
  - Soldier Patient Tracker/Location
  - Warrior Transition (WT)
- Health Professional Education, Professional Development , and Special Pay
  - Continuing Medical Education (CME)
  - Medical Education (MEDEd)
  - Health Professions Scholarship Program (HPSP)
  - Health Care Specialist Tracking System 68W
  - Special Operations Forces (SOF)
- Health Command and Decision
  - Command Management System (CMS)

The intent of a data model is to identify and define the business perspectives with information use in mind.

The Conceptual Model (see Figure 10) captures data categories and major data relationships that reflect high-level data scope. It also supports understanding and discussion with stakeholders in a technology independent way. The Conceptual Model can usually be drawn using standard desktop tools (PowerPoint, Visio).

The Logical Model, on the other hand, captures a full enumeration of data content for the DW from a business perspective, also technology independent. It includes both relational and dimensional schemas, and is designed using data design tools (e.g., ERwin).

The Physical Model is a technology dependent model (i.e., database/platform-specific). It is also designed using data design tools (e.g., ERwin).



Conceptual Model Example (notional)

- Analyzing the scope of MODS vision will produce the major categories of data and their key relationships
- Conceptual model would be refined by analyzing the information needs (e.g., BQs and measures)

Soldier Medical Readiness, Non-Readiness & Surveillance
Health Professional Education, Professional Development and Special Pay
Health Command and Decision Support
*Soldier (Person) Demographic and Organizational Data
Training / Certification
Earned by
Soldier
Assigned
Assignment
Possessed by
Skill
Belongs to
Characterized by
Unit
Characterized by
Readiness Status
Etc.

Figure 10. Conceptual Model Example

C.7 Known Data Quality Issues and Considerations

The following issues were discussed and defined for the current MODS DW environment:

Table 1. MEDCOM MODS Data Quality Issues

Data Quality Issue	Data Quality Factors
1. **Soldier Readiness Processing (SRP) events.** In order to link various events (e.g., Deployment Health Assessments (DHAs), profile reports, etc.) to an overall SRP event, today's reports must make certain timing assumptions that will relate them. Going forward, a similar approach will be in need within the MODS DW.	Data consistency
2. **MEDPROS Web vs. Mainframe.** Given the timing differences as to when each system is updated with information from each other, there can be a one-day lag in the information, making reports inconsistent with each other. Further, reporting out of either system could produce different results throughout the	Data consistency Data timeliness

day, as updates are made. While each system can be updated in near-real time, they don’t update each other except once per day. Since many updates could have been made to the data within one day, making the delta between them significant (e.g., as much as 10K records or more could be out of sync). This could greatly impact performance measurement reporting and decisions made on the data. Going forward, the intent is to stage within the MODS DW a snapshot for reporting purposes that is of a set and known timeliness. Currently, the requirement is for a one-day lag in the data.	
3. **Warrior Transition (WT) Information.** There is not a full snapshot of those in WT (e.g., those assigned, attached, or in-transit), thus WT information may be inconsistent. This is a result of medical readiness’ lack of a full view into a soldiers’ status. And there may be differences in UICs, such as their home UIC vs. their deployed, WT, or attached UICs.	Data consistency Data completeness
4. **Pregnancy information.** To date there has been a problem with the accuracy of pregnancy information, with records showing the condition for more than a year in some cases (as many as 6-8K records). This information is expected to improve with the use of the eProfile system.	Data accuracy
5. **Provider Assignment information.** Currently, knowing the providers who have been assigned to health assessments is not readily available and is not always captured within source data. This issue relates to the data categories for both *Provider Assignment* (e.g., event relationships such as DHAs) and *Medical Staff Assignment* (e.g., locational assignments, such as “who are the providers at Fort Hood?”). While Provider Assignment information should be available in DHAs (e.g., through a digital signature, Army Knowledge Online), that information is not currently accessible to MEDCOM users.	Data availability
6. **Immunization information.** There is some duplication of immunization records in MEDPROS due to reference on dates of shots given. While the Defense Enrollment Eligibility Reporting System feeds this information to MEDPROS, MEDPROS also supports direct input of this information. Further, there is incomplete shot information, such as manufacturer and lot #, which if provided could help to reduce the duplication problem.	Data non-redundancy Data completeness
7. **Force Structure information.** Perhaps one of the biggest challenges has been to maintain what are essentially many different organizational hierarchies, depending upon the perspective and use of the information. The views needed for both the commanders’ and Defense Medical Information System (perspectives are considered to be fairly accurate. However, the Force Management Support Agency hierarchy is considered to be accurate down to the company level (for all three major components). Unit information below company is managed/defined by the units themselves and does not make its way into formal systems. This is especially true for those in a war fighter status. A new effort, called the Global Force Management Initiative, is expected to address these gaps in the future.	Data accuracy Data completeness

8. **UIC information within HR/TAPDB.** Currently, there can be significant delays between receiving updated UICs for soldiers. While a unit that is "losing" a soldier to another unit (the "gaining" unit), can update *their* records accordingly, the soldier is not officially "moved" to the gaining organization within the HR/TAPDB system until a personnel action has been posted by the gaining organization. It was the view of the participants that the DW will be unable to address this problem directly.	Data consistency
9. **Lab information from other military branches.** There are timing issues with information from the Air Force and Navy due to their processing differences. While the Navy uses the same lab as the Army, due to release processing delays, it can often take 30-60 days longer to receive their records than that of the Army. For the Air Force, which uses a different lab than the Army and Navy, their data is typically sent as a few batches per year. Also, the Air Force does not use MEDPROS to capture all of its readiness information.	Data timeliness
10. **Deployment Event information.** High-level deployment information is available post-deployment (i.e., retrospectively). Deployment operation, theater, country, and periods (to the month-level) are typically provided on the Post-Deployment Health Reassessment Program Reports. Regardless, MEDCOM would benefit from using this information to assess readiness trends as related to various deployment scenarios.	Data timeliness Data precision