WHITE PAPER

Data Repository Guide

CompTIA

Introduction

Under the Space Policy Directive 3 (SPD3), the National Space Traffic Management Policy (issued on June 18, 2018), the Department of Commerce was called upon to *"be responsible for the publicly releasable portion of the DOD catalog and for administering an open architecture data repository"*. According to SPD3, the essential features of the open architecture data repository will include:

- Data integrity measures to ensure data accuracy and availability;
- Data standards to ensure sufficient quality from diverse sources;
- Measures to safeguard proprietary or sensitive data, including national security information;
- The inclusion of satellite owner-operator ephemerides to inform orbital location and planned maneuvers; and
- Standardized formats to enable development of applications to leverage the data.

The Department of Commerce goals for a "state of the art" data repository

The current focus is on satellite sensor data and accompanying analytics. The Commerce repository will need to be based on a framework that helps develop and guide international standards and best practices. To date, there are over 50 organizations that are creating space safety standards; there needs to be unity. By leveraging NIST, there will be credibility on standards and best practices. The goal of the data environment is to have an open environment to facilitate creating, developing, testing, and validating against the data on an on-going basis for the public, commercial, and international stakeholders. The architecture will need

to integrate DOD data (with a focus on interoperability, which is the ability of computer systems or software to exchange and make use of information) and be more extensive than space-track.org. Commerce will initially look at rapid experimentation focused on sensor data and cyber … international is likely further out. Within the repository, Commerce will need a mechanism to ensure data quality, attributes, pedigree of data, and phenomenology.

CompTIA recently created the Blockchain guide "Harnessing the Blockchain Revolution: CompTIA's Practical Guide for the Public Sector". The Guide takes a holistic approach to examining the Blockchain ecosystem, focusing on benefits, challenges and security issues, applications by use case and sector, federal/state/local Blockchain programs, adoption considerations and policy recommendations. We will take a similar approach for our Data Repository Guide.

What is a data repository?

A data repository is a manageable collection of databases with corresponding metadata which allows for the storing and sharing of data. One of the most relevant characteristics of a data repository is that it has the capability to be searched and mined across an immense swath of data. The mined data can then be analyzed and provide significant potential efficiencies across various domains.

To date, data repositories have primarily been used within the scientific community. However, times are changing. Former Google Chairman Eric Schmidt recently stated that "we produce more data every other day than from the beginning of civilization until the year 2003 combined". Data is increasingly being tied to, and helping drive decisions for, an entity's mission. With the proliferation of data, there will be a pent-up demand for centralized locations to access large amounts of data.

In the space community, data repositories have historically been stove-piped. The construct of the Commerce data repository needs to be fundamentally different as the Repository will conceivably have millions of data points. According to the European Space Agency, there are 29 000 objects larger than 10 cm, 750 000 objects ranging from 1 cm to 10 cm, and 166 million objects from 1 mm to 1 cm. The 166 million object figure is probably are underestimation as miniscule fragments are nearly impossible to track. According to a 2011 National Research Council report, the amount of debris has reached a tipping point, with enough currently in orbit to continually collide and create even more debris, raising the risk of spacecraft failures.

In our Guide, we will discuss the benefits of a collaborative repository, discuss technology and architecture options, examine some of the best practices specifically from the satellite and scientific communities and then relevant implementations from the public and private sector.

Benefits of Data Repository Use

A robust data repository can bring a range of benefits to a community to improve both collaboration and commerce. In the ever-changing world of new processing and storage technology, the benefits of integrating datasets and databases in a central location are more prominent than ever. While there are numerous benefits to the creation and use of a community data repository, we will focus on three specific benefits aligning with the Department of Commerce's mission.

Increased Data Quality Leads to Faster, More Accurate Decision-making

Having the right information on hand at the right time can influence key business decisions and drive the prioritization of ongoing initiatives. When dealing with materially sensitive and expensive items that assist with common public utilities like GPS, having accurate, high quality data is extremely important. This need carries over into the space community, since accurate object positions and orientations are critical to interpreting the potential object conjunctions as well as the sensor data collected from satellites. How does a data repository ensure that the data is of high quality, high integrity? According to a study on The Economics of Data Integrity by the Online Computer Library (OCLC), "*... ensuring the* quality *of data requires specific subject area expertise, due to varying needs of the disciplines. This is perhaps an even more significant opportunity for economies of scale. If every repository had to have a wide range of subject experts, the costs would be prohibitive.*" Ensuring that all data is properly described semantically by the dataset owners ensures a more accurate representation, and allows cross-comparisons between datasets that have the same data – identifying areas that need cleansing or improved quality.

Preserving Historical Data Assists with the Valuation of Goods & Services

The value of datasets is often derived from recent data, such as recent satellite images that provide the data to analyze current conditions. Historic datasets also provide significant value for trend analysis, for the study of potential drift in baseline measurements, or for indications of changes in data collection methods. With a central repository integrating information and the right amount of storage, historical data is easily retained and applications accessing that data can be upgraded or modified seamlessly. Perhaps more importantly, preserving historical data also leads to more insights, which can also mean more economic value. For example, maintaining historical cost data for a single component of a satellite will lead to better forecasting for that component.

A Central Data Repository Enhances Data Sharing & Promotes Collaboration

We've already stated that integrating datasets through semantic metadata increases the quality and integrity of the data because multiple stakeholders are accessing the same logical model. The centralization of data representation also leads to collaboration and data sharing for that same reason. It is important that audit trails are maintained within a repository so changes can be tracked and all users of that data know whom to contact if something seems awry. This not only builds in organic validation of the data; it also encourages cross-organization collaboration and breaks down silos. Users of the same data have the means to connect with other users, promoting a holistic, data-driven culture.

Aforementioned benefits are just a few of the reasons why a centrally integrated Data Repository is the right solution for Commerce's data challenges. Such a semantic Data Repository would be beneficial to solve many common data problems including siloed data, inconsistent or incorrect data, and can provide a "network effect" of value in that the more data is integrated, the more value it would bring to the space commerce, operations, and research communities.

Data Repository Cheat Sheet

Characteristics	Relational Database	Data Warehouse	Data Lake	Data Mart	Operational Data Store
Data Types	Structured, numerical data, text and dates organized in a relational model	Relational data from transactional systems, operational databases and applications	Structured and unstructured data from sensors, websites, business apps, mobile apps, etc.	Relational data subsets for specific applications	Transactional data from multiple sources
Purpose	Transactional Processing	Data stored for business intelligence, batch reporting and data visualization	Big data analytics, machine learning, predictive analytics and data discovery	Data used by a specific user community for analytics	Ingest, integrate, store and prep data for operations or analytics; often feeds a data warehouse
Data Capture	Data captured from a single source, such as a transactional system	Data captured from multiple relational sources	Data captured from multiple sources that contain various forms of data	Data typically captured from a data warehouse, but can also be from operational systems and external sources	Data captured from multiple enterprise applications/ sources
Data Normalization	Uses normalized, static schemas	Denormalized schemas; schema-on-write	Denormalized; schema-on-read	Normalized or denormalized	Denormalized
Benefits	Provides consistent data for critical business applications	Historical data from many sources stored in one place; data is classified with user in mind for accessibility	Data in its native format from diverse sources gives data scientists flexibility in analysis and model development	Easy, fast access to relevant data for specific applications and types of users	Fast queries on smaller amounts of real-time or near-real-time for reporting and operational decisions
Data Quality	Data is organized and consistent	Curated data that is centralized and ready for use in BI and analytics	Raw data that may or may not be curated for use	Highly curated data	Data is cleansed and compliant, but may not be as consistent as in a data warehouse

Source: TechTarget

Data Repository Technology and Architecture

The technologies for handling large volumes of data fundamentally changed in the mid-2000s. This was the point in time where mainstream data-intensive systems were *parallelized*. Previously "bigger" data was handled through 'vertical' scaling, with faster processors and increased hard-drive capacity. After the big data paradigm shift, data was distributed 'horizontally' across multiple nodes for scalable, efficient, cost-effective processing. This shift to parallelization in data-intensive processing is analogous to the shift decades ago to parallelization in compute-intensive systems – a field we call High Performance Computing. Big data systems can be described in terms of the infrastructure, storage platform, and processing frameworks (see NIST 1500-6 reference architecture).

Infrastructure for big data systems can leverage *cloud*, or they can be established on the *bare-metal* or physical hardware systems. While cloud infrastructures do add additional tools into the mix (such as Amazon Web Services Redshift), other tools can run equally well in the cloud or in on-premise resources.

The shift in *storage platforms* began with the expanded file systems to handle larger block sizes through Hadoop Distributed File System, then several non-relational databases (known as *NoSQL*) were introduced that allowed data to be spread across several data nodes to be processed in parallel. These databases sacrificed some of the properties of relational databases to gain speed – by distributing portions of the data across several data nodes for independent processing (redundant sentence?). The initial types of NoSQL databases were key-value, document, big table, and graph. Subsequently, additional techniques emerged to restore some of the relational database properties to distributed data systems, and these began to be termed NewSQL.

The accompanying change in data *processing* began with the MapReduce (M-R) technique. In M-R, the same data processing query was 'scattered' to each of these many separate data nodes, and the results were 'gathered' back to a single compute node. There were some inefficiencies in how the tasks were executed, leading to the introduction of MapReduce 2.0, known as YARN, which improved cluster resource management. In the subsequent years,

analytic frameworks such as Spark were introduced to insulate the data science applications from the lower-level implementation details.

The *architecture* of a big data system depends on the characteristics of the dataset, including volume, velocity, variety, and variability. The analytics lifecycle can be described as *collect*, *curate* (cleanse, organize,), *analyze* and then *act* on the resulting knowledge. Traditional *data warehouses* collect the data to temporary storage, curate (through ETL) and store the cleansed data into a warehouse, then analyze the data and prepare it for some action by the organization. By contrast, large *volume* data systems store the data as it's received. There is too much data to move, so any cleansing and analytics are done together dynamically. For high *velocity* data – previously known as data streaming – the data is analyzed dynamically (typically using in-memory systems) and potentially some data summaries are stored after the analytics or alerting tasks. *Variety* data adds the additional complication of having very different datasets that need to be integrated dynamically, perhaps from datasets that the organization does not own. *Variability* refers to datasets that have changing characteristics. From an architecture point of view this implies that the analytic system may need more or fewer resources. In this kind of "bursty" data processing, it can be cost-effective to use cloud systems where you can dynamically add or remove compute resources.

Taken together, the choices of specific tools and techniques for infrastructure, storage platform, and processing make up an organization's enterprise data repository. The use of NoSQL systems in the storage platform has led to the term *data lake*, where multiple datasets are stored in various NoSQL databases in this data lake, then work begins on integrating and processing the data that is needed for the application. A data lake typically implies that the organization is housing and controlling all their data – a centralized model.

Repositories for Collaboration

The model for allowing analytics across databases that are disperse is quite different from the traditional *data warehouse* or new data lake. This de-centralized control of data occurs when an organization needs to leverage data that it doesn't own, or when a community wishes to share their data and collaborate. In the parlance of big data, this is a *variety* scenario. In a decentralized model some datasets may reside within the organization, but others may be housed in other databases, other data lakes, even in other organizations outside your control. A collaborative repository needs to strike a balance between the community data to be ingested and controlled centrally, versus the data that will be managed and governed separately by other organizations yet be easily accessible by the collaborators.

In traditional relational databases, providing an over-arching structure that allows integration of disparate distinct databases is known as a *federated database*. Such databases require the same rigidity of data integration as data warehouses, they are just separate databases. With the advent of NoSQL databases or data lakes, the schema - or organization - is not imposed prior to storage but is implemented as the data is retrieved for analysis. Consequently, instead of storing the data in a highly structured way, we instead map the data to a highly structured semantic representation for interoperability.

Gartner describes the need to develop information sharing environments to develop data economies. *"In the data economy, effective information sharing is vital to reaching customers, opening markets, improving transparency, sparking innovation and improving trust among stakeholders."* [Gartner G00213344]. While initially conceived to rescue enterprise data from within data siloes, the Gartner Logical Data Warehouse (LDW) concept can apply equally to community repositories as well. The LDW is essentially a clearinghouse for information where the data resides in many organizations or systems. This has the advantage of allowing the individual groups in the collaboration/ organization to control their own data; advantageous especially when the new datasets are included or when the data is proprietary and not freely available.

What is new in an LDW is that the data integration must be performed automatically; meaning the data in each repository must be described semantically – according to what the data refers to. Secondly the stakeholders must be able to find the data according to what it represents, not where it is physically located or who controls it.

Implementing the vision of an LDW requires the use of semantic definitions, ideally organized through an ontology. An ontology allows for the categorization of information and the encoding of precise descriptions of what the data represents and how it relates to other data. The community collaborators can then select the data they want from this semantic metadata catalog. For example, a query could request satellite images taken at a given set of frequencies in the spectrum, for this spot on the earth's surface, over a specific timespan. Only if the community stakeholders can gather the data they want semantically

– rather than having to know all the places where the relevant data might reside - will a community flourish. The technical magic is to then create the middleware to translate this metadata selection into queries against the databases, tables, and elements where the data resides.

DARPA has been developing the Hallmark ontology for the precise description of objects and their locations in space. This effort can form the foundation for the semantic organization, data catalog, and data retrieval mechanisms for military, scientific, and commercial datasets valuable to growing space commerce. NASA Langley has also worked on an ontology for science-oriented data retrieval among all the atmospheric measurements and model results that they have in their repository.

Security

Organizations are under continual cyber-attack, the eventual end target typically being the exfiltration of data, the deletion of data, or the deliberate alteration of data for operational or economic harm. For proprietary data, the concern is focused on ensuring that unauthorized users are not able to see the data. This is typically done through a security-in-depth approach that ends in the encryption of data elements in case all other controls are breached. The specific concerns for an open repository, one that makes data freely available, focus on ensuring that the backend systems are not accessible from the open internet except through a strict role-based access control. The implication is two-fold. The dataset themselves need to be tagged (with semantic metadata as discussed above), and the stakeholder roles need to be equally tagged with the types of data they can read, write, or update.

In a public-serving repository to promote community safety and economic growth, a range of stakeholders need to interact with the repository to gain situational awareness. Partners will have write access to contribute new data to their own datasets through APIs – which must ensure the data is properly tagged. Stakeholders such as down-stream websites that would retrieve datasets must have APIs that allow them to query through the semantic catalog to get to the appropriate data no matter where it resides. The public will need read-only access controls that limit their access to the freely available portions of the data. In some cases, the data may be proprietary, and a monetary structure will need to be included in the metadata to ensure the consumer is aware of the availability of the data but is also

aware of the monetary cost. Each of these data and user roles will need be aligned and enforced to ensure proper usage of the data; whether it is held in a central repository or whether the catalog points to a database outside the organization housing the central repository.

Some repositories, such as those involving space objects, have additional security concerns involving classified information. In the case of tracking objects in space, some objects are publicly described, and the sensor data they collect made freely or commercially available. Other objects belong to the military and are classified. While their position, size and orbits are of significant interest to others, to be forewarned of potential collisions, the precise orbit, specific configuration, and sensor characteristics may remain classified. Data will be contributed to the unclassified repository based on information that can be released from classified systems. While the collision reports from the open community repository would be available to all, the burden falls upon the classified systems to confirm the potential of a collision, and to provide a mitigation request back through the repository to the operation of the other object.

Security and Culture

Often overlooked, culture plays a critical role in being able to operationalize a multi-player/organization data repository. Working across many verticals (civilian, military, commercial, public) who each have very distinct cultures can be a monumental task. Cultural creep can set it and inadvertently expose one to cyber-attacks. Expectations should be clearly spelled out from the very start to help offset any cultural bias. A potential solution to security and culture is to follow the soon to be offered FedRAMP training from GSA. This training would be available to all government officials. The concept would be like the training program offered by the Defense Innovation Unit named HACQer.

FedRAMP: Security Requirements for Consideration

The Federal Risk and Authorization Management Program (FedRAMP) is a government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services. The following sections from the FedRAMP Security Assessment Framework (SAF) outline a prudent approach for the Commerce Data Repository security requirements.

In the document phase of the SAF, Steps 1-3 of the Risk Management Framework will be covered by categorizing the information system, selecting the security controls, and implementing and documenting the security controls and implementations in the System Security Plan (SSP) and supporting documents.

3.1.1 CATEGORIZE THE INFORMATION SYSTEM

To categorize the system, the CSP determines the information types and completes a FIPS PUB 199 worksheet to categorize what types of data are (or can be) contained within the system to determine the impact level for the system. The categorization is based upon NIST Special Publication 800-60 (Volumes I and II) Guide for Mapping Types of Information and Information Systems to Security Categories.

The analysis of the data contained in the system, based upon the information in the FIPS PUB 199 worksheet, will determine if the security categorization for the system is at the Low, Moderate, or High impact level. Currently, FedRAMP only supports security assessments of systems that have Low or Moderate impact levels. A template for the FIPS PUB 199 is available on www.fedramp.gov

3.1.2 SELECT SECURITY CONTROLS

After completing a categorization in accordance with FIPS PUB 199, the CSP selects the FedRAMP security controls baseline that matches the FIPS PUB199 categorization level from Section 3.1. The FedRAMP security control baseline is published on www.fedramp.gov. Additionally, Section 13 of the FedRAMP System Security Plan Template summarizes the controls for both Low and Moderate security impact level systems.
The FedRAMP security control baseline provides the minimum set of controls that CSPs will need to implement to meet FedRAMP's requirements for Low or Moderate security impact level systems.

3.1.3 IMPLEMENT SECURITY CONTROLS

Once the CSP has selected the FedRAMP security control baseline, the next step is to implement the

security controls related to that impact level. For most providers, many of the controls are already implemented but need to be described adequately within the FedRAMP templates. Some controls might require the implementation of new capabilities, and some controls might require a re-configuration of existing implementations.

The FedRAMP program considers that systems may vary between vendors and allows some flexibility in implementing compensating controls or alternative implementations. The imperative part of implementing security controls is that the intent of a security control is met. CSPs may provide alternative implementations that demonstrate the implementation satisfies the intent of the control requirement. For any control that cannot be met, CSPs must provide justification for not being able to implement the control.

3.1.3.1 SYSTEM SECURITY PLAN

After implementing security controls, CSPs must document the details of the implementation in a System Security Plan. Every security package must include an SSP based on the FedRAMP template. All cloud providers must use the FedRAMP template, regardless of what type of ATO they are vying for. The SSP describes the security authorization boundary, how the implementation addresses each required control, roles and responsibilities, and expected behavior of individuals with system access. Additionally, the SSP allows AOs and review teams to understand how the system is architected, what the system boundaries are, and what the supporting infrastructure for the system looks like.

The SSP template can be found on www.fedramp.gov. Additional guidance about how to describe control implementations in the SSP can be found within the SSP template.

3.1.3.2 INHERITING CONTROLS FROM A LOWER-LEVEL SYSTEM

In the cloud space, many cloud systems rely on other cloud systems to provide a comprehensive set of services for the end customer. An example of this is a software provider utilizing an infrastructure provider to deliver the Software as a Service (SaaS). In this case, the software provider will inherit security controls from the infrastructure provider.

The FedRAMP SSP template provides for marking a control as inherited and from which system that control is being inherited. By allowing for inherited controls, FedRAMP enables the stacking of authorization packages like building blocks. In this model, the SSP for each system must only describe the implementation of that specific system (for example, SaaS service providers in the example above would not detail any implementation details of the leveraging infrastructure provider within the SaaS service SSP). This eliminates redundancy across authorization packages and keeps authorizations delineated by system.

Much in the same way the software provider in the example above relies on the infrastructure provider to deliver services, the software provider also relies on the security implementations and authorization of the infrastructure provider for the software provider's implementations and authorization. Accordingly, if a CSP has inherited controls within the System Security Plan, the authorization of that CSP will be dependent on the authorization of the CSP whose controls they inherit and systems they use to deliver the end service.

Case Studies and Best Practices

Satellite Industry Data Repositories

Alaska Satellite Facility SAR Data Center

www.asf.alaska.edu/

The SAR Data Center has a large data archive of Synthetic Aperture Radar (SAR) from a variety of sensors available at no cost. Much of the SAR data in the ASF SDC archive is limited in distribution to the scientific research community and U.S. Government Agencies. In accordance with the Memoranda of Understanding (MOU) between the relevant flight agencies (CSA, ESA, JAXA) and the U.S. State Department, the ASF SDC does not distribute SAR data for commercial use. The research community can access the data (ERS-1, ERS-2, JERS-1, RADARSAT-1, and ALOS PALSAR) via a brief proposal process.

Central Satellite Data Repository Supporting Research and Development

www.star.nesdis.noaa.gov/star/index.php

Near real-time satellite data is critical to many research and development activities of atmosphere, land, and ocean processes. Acquiring and managing huge volumes of satellite data without (or with less) latency in an organization is always a challenge in the big data age. An organization level data repository is a practical solution to meeting this challenge. The STAR (Center for Satellite Applications and Research of NOAA) Central Data Repository (SCDR) is a scalable, stable, and reliable repository to acquire, manipulate, and disseminate various types of satellite data in an effective and efficient manner. SCDR collects more than 200 data products, which are commonly used by multiple groups in STAR, from NOAA, GOES, Metop, Suomi NPP, Sentinel, Himawari, and other satellites. The processes of acquisition, recording, retrieval, organization, and dissemination are performed in parallel. Multiple data access interfaces, like FTP, FTPS, HTTP, HTTPS, and RESTful, are supported in the SCDR to obtain satellite data from their providers through high speed internet. The original satellite data in various raster formats can be parsed in the respective adapter to retrieve data information. The data information is ingested to the corresponding partitioned tables in the central database. All files are distributed equally on the Network File System (NFS) disks to balance the disk load. SCDR provides consistent interfaces (including Perl utility, portal, and RESTful Web service) to locate files of interest easily and quickly and access them directly by over 200 compute servers via NFS. SCDR greatly improves collection and integration of near real-time satellite data, addresses satellite data requirements of scientists and researchers, and facilitates their primary research and development activities.

NORMAP: Norwegian Satellite Earth Observation Database for Marine and Polar Research

www.nersc.no/project/normap

The Norwegian Satellite Earth Observation Database for Marine and Polar Research (NORMAP) infrastructure project was established to help create a data repository, including a meta database, for Nordic and Arctic regions for Earth observation data from polar orbiting satellites to facilitate and stimulate high quality and original multidisciplinary Earth System research and application

in marine, polar and climate sciences. Within the first two years, NORMAP seeks to make available a set of selected quality controlled multidisciplinary scientific data products that will support air-sea-ice process studies, near real time applications and long time series for climate change purposes. Longer term, a goal (among others) is to advance the effective use of satellite EO data by the scientific community so less time is spent searching and qualifying data giving more time to scientific studies and analysis.

Geospatial Cloud Analytics (GCA)- DARPA

www.darpa.mil/program/geospatial-cloud-analytics

The Geospatial Cloud Analytics (GCA) program is developing technology to rapidly access the most up-to-date commercial and open-source satellite imagery as well as automated machine learning tools to analyze this data. Current approaches to geospatial analysis are ad hoc and time intensive, as they require gathering and curating data from many available sources, downloading the data to specific locations, and running it through separate suites of analytics tools.

GCA aims to virtually aggregate vast amounts of commercial and open-source satellite data that is available in multiple modes—optical, synthetic aperture radar (SAR), and radio frequency (RF)—in a common cloud-based repository with automated curation tools. The platform and tools would provide DoD geospatial analysts global situational awareness, event detection, monitoring, and tracking capabilities beneficial to U.S. forces around the world.

In addition to developing a scalable geospatial data platform with tools and a user interface, GCA aims to create analytical applications that would allow analysts at the operational and tactical level to draw specific information from the aggregated data. GCA will pilot an analytical services business model where commercial entities offer analytics services and apps via a competitive marketplace.

Case Studies from Industry Leaders

Data Repository for a Federal Sector Organization

An organization in the federal sector was unable to access critical, cross-business line information for executives, managers, and external stakeholders in a timely and accurate manner. Additionally, extracting information from many disparate data systems required considerable investments of time and resources and, when the processes used to accomplish these goals was replicated, it failed to provide consistent or repeatable results. A lack of data consistency and integrity across multiple sources has led to skepticism about the quality of the information and the ability to report on activities with assurance.

The solution was to create a central data repository in the form of an Enterprise Data Warehouse (EDW). The EDW consists of multiple warehousing data applications and business intelligence tools, providing thousands of reports to multiple users worldwide, directly and securely to their desktops, via the web. Data is obtained from the source systems using various mechanisms. The source systems are made up of different database types and the data is extracted on a routine basis so it is always current.

Data Architecture and Warehouse Solution

Each year, an organization receives several hundred data and report requests on matters of accounting, budgeting, procurement, logistics, financial systems, policy, planning, and audit oversight that support mission critical decision-making across the enterprise. Responding to these requests requires a high level of effort for several reasons, such as:

- Data needed to satisfy the requests can be stored in multiple, siloed sources
- Data needed to satisfy the requests can be stored as structured or unstructured data

- Assembling data to satisfy the requests is a manual exercise, having the potential of being inconsistently rendered, even for identical data requests
- Exact or near-duplicates requests require the same amount of effort to satisfy each
- Data quality errors are detected only if noticed by the assigned staff

As a solution, the organization was able to leverage existing business and technical infrastructure to design and implement changes in data management that improve their ability to service requests for data and information from customers within and outside the business. The key element in the solution was the design and development of a data warehouse, including a selection of relational and alternative architectures to support storage of structured and unstructured data. From there, the organization was able to produce real-time data visualization, business intelligence, reporting, and analytics to quickly manage and respond to report requests.

Federal Government

The National Cancer Institute (NCI) - as well as other programs of the National Institutes of Health – have developed controlled vocabularies for communication and collaboration amongst scientists and clinicians. This ensures that information from a wide array of fields use the same terminology, or terminology that can be translated between disciplines. This was accomplished through design and ongoing implementation of a program of Common Data Element Harmonization in the Cancer Data Standards Repository (caDSR). This harmonization required an understanding of the data that supports clinical trials and cancer genomics, application of best practices in metadata creation and management, collaboration with a large group of stakeholders from many cancer centers, and a staff of experts to support research on the cause, diagnosis, prevention, and treatment of cancer.

Intelligence

The defense intelligence community has implemented an ontology-based data integration called the Object Based Production (OBP). The OBP methodology defines the relationships between objects/entities, and points to the original data source (documents or structured databases). The ontology-based organizes the information from many sources by linking the information details to real-world objects and presents the information to an analyst in the terms that are used for operations – such Facilities, Equipment, Organization, People and Activities and Events. Enriched objects are presented to analysts via visualization layers, or to any analytic tool needed for further intelligence enrichment.

USDA FS (US Department of Agriculture - Forest Services)

The Forest Services (FS) developed an enterprise data warehouse (EDW) as part of an Information Management program. This program has developed of an EDW suite of applications and toolsets, continued to expand the EDW content, promote EDW use through outreach and training, and support installation and implementation of the tools. As a result of these efforts, the FS created an enterprise business intelligence environment to serve as an official source of trusted information drawn from authoritative source systems. The EDW provides a means of integrating data from many stovepipe applications and data warehouses, formatting it for ease of use, and making it available via multiple mechanisms. These mechanisms include web services, map services, downloadable data files, direct access via ArcGIS, and Cognos reports. EDW data is now published on the FS Geodata Clearinghouse and data.gov and supports a growing user community.

Military

The Joint Forces Combatant Command (JFCC) SPACE provides Space Situational Awareness (SSA) data sharing for military, civilian, commercial, and foreign customers through the space-track.org website. The website provides data on 2200 on-orbit payloads, provides conjunction data and analysis to 404 organizations of 1200 users, and supports 147,000 hobbyists and academics. This project maintains and improves the website, screens submitted data for compliance with data standards, and supports rapid communication among users. The site can dynamically change the data displayed on the web pages by user, group, and role; for example, revealing redacted Conjunction Data Message data only to authorized personnel. The number of queries is limited, and each is analyzed to restrict queries that would overwhelm the database. Space-track segregates spaceflight safety data from non-critical data to ensure critical traffic is not slowed down. They maintain the system on virtual machines to improve backup and restore capability, as well as quickly mitigate any potential spillage from the information they receive from classified systems. The site minimizes its attack surface by proactive monitoring industry notifications and application of patches to remove system vulnerabilities, not waiting on notification

through formal channels. This project follows an agile development methodology to accommodate customer needs and the incorporation of new datasets and uses Splunk for site logging and auditing to diagnose bandwidth usage spikes and spot users who may be abusing their website agreement.

State and Local

A consortium of federal agencies developed the National Information Exchange Model (NIEM), a law enforcement Information Sharing Framework that brings communities together, builds consensus, and works toward a better government through the sharing of data. Through a set of common, well-defined data elements for data exchange development and harmonization across agencies, the Department of Homeland Security (DHS), Dept. of Justice (DOJ), and Department of Defense (DoD) have enhanced their collaborative partnership with agencies and organizations across all levels of government. NIEM has been recognized as the "Standard for Information Sharing". NIEM exchanges are being developed to allow for a more efficient and consistent method of sharing important information, representing a significant first step in the development of a borderless network of information exchange between the U.S., Canada, Europe and Mexico.

Commercial

The use of clinical data in healthcare- Texas Children's Hospital Case Study

www.healthcatalyst.com/news/teams-and-tools-unlock-millions-in-savings-at-texas-childrens/

Texas Children's Hospital began work on an enterprise-wide data warehouse solution to implement a clinical, analytical and process-based framework. IT staff and clinicians started working in cross-functional teams, using new analytics applications that enabled them to better visualize results in minutes, opposed to months. With data emerging from different systems within the hospital, researchers also faced limitations conducting analysis.

The organization realized it needed something beyond the records system if it was going to succeed in using the data on its patients to achieve quality improvement The enterprise-wide data warehouse is a large step toward creating a repository of information that enables a consistent view of data from many sources within the organization. That, as a basis, can help bolster research findings that can be backed up by the actual experience of the organization.

Texas Children's executives estimate they have achieved about $4.5 million of direct benefits from only four of its EDW projects. Davis said the organization is benefiting as the warehouse can make use of meaningful data that previously had been "trapped" in the EHR.

Architecture and Design Recommendations

For CompTIA Data Repository we recommend adhering to the following overarching principles in solution design.

Stakeholder Requirements

Ensure that all stakeholders are identified, and their needs translated into requirements. This will ensure that security, classified versus unclassified, operations versus research, proprietary versus open, and commercial versus government requirements will all be captured.

Flexibility

Apply a modular distributable architecture, which isolates the complexity of integration, business logic, and persistence from each other to enable the easy integration of new technologies and processes within the application.

Standards-Base

Comply with established industry standards, where possible. Apply lists of values for data entry where possible, using standards like NIEM to increase interoperability. Standardized data should also be defined specific to the business and published at the enterprise level. The standards compliance will not only apply to development but also to design, platform/ infrastructure, and other parts of the repository.

Scalability

Ensure that design enhancements can scale horizontally to enable handling increasing numbers of collaborators, datasets, and external databases. Explore the differing tiers of storage (long term, disk, memory) to both meet data retrieval timeliness requirements while maintaining cost-effectiveness.

Platform Independence for Consumers

Make sure to achieve technology platform independence by leveraging web services so interoperability will be achieved through the data integration, not tightly coupled systems integration.

Taxonomy development

Develop an ontology to allow seamless integration of disparate datasets. Each new dataset only has to represent itself according to the ontology, rather than trying to integrate point-to-point with all other datasets. This allows the ontology to serve as the table of contents for all the datasets – allowing not only data selection by its content, but also automated data fusion for cross-dataset queries.

References

NIST SP 1500-x, 9 volumes in the Big Data series (https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-10.pdf)

Gartner LDW
https://blogs.gartner.com/henry-cook/2017/05/23/building-the-logical-data-warehouse-ldw/

Digital Guardian- What is a Data Repository?
https://digitalguardian.com/blog/what-data-repository

BU Data Services- What is a Data Repository?
https://www.bu.edu/data/share/selecting-a-data-repository/

MSU 2016- The roles of a data repository
What should a data repository do?

NIST Science Data Portal
https://data.nist.gov/sdp/#/

FedRAMP
www.fedramp.gov

Glossary

Data repository- is a manageable warehouse of databases with corresponding metadata which allows for the storing and sharing of data.

Data lake- a centralized repository that stores both structured and unstructured data.

Data warehouse- the operational system within which value-added services such as data analysis resides.

Decision aid- tools that are utilized for shared decision-making.

Unified Data Library (UDL)- fragmented data sources that are merged into one, single central view.

API- the operational aspects that allow the creation of applications.

Metadata- data that provides a description and context of the data.

Cloud computing- the practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer

Hybrid Cloud- is a combination of on premises, private, and public cloud services.

FedRAMP- The Federal Risk and Authorization Management Program (FedRAMP) is a government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services.

Space Situational Awareness- in depth and detailed knowledge of the near space operational environment.

Volume- the principal characteristic that makes data "big".

Variety- the compilation of structured and unstructured data.

Veracity- trustworthiness of the data.

Velocity- frequency of the data streams.

Acknowledgements

CompTIA would like to thank the important contributions to this guide made by the following organizations:

CompTIA.org