The DATA Act – IT Infrastructure Guidance

Change Facilitation for IT Departments

Collaboration & Transformation (C&T) Shared Interest Group (SIG)
Financial Management Committee
DATA Act – Transparency in Federal Financials Project
Date Released: August 2015

SYNOPSIS
Planning and implementing the IT changes required to meet the requirements and objectives of the DATA Act (P.L. 113-101), while continuing to meet mission and business goals, is a both a significant challenge and opportunity for engaged IT departments. This report outlines key guidance for IT strategy and operations to facilitate the changes required, while continuing to advance IT modernization and simplification progress. Key areas to address that can set the stage for success – and perhaps even catalyze it - include IT infrastructure consolidation, engaged data governance, and virtualization. This paper explores each of these IT infrastructure governance and investment areas, as pragmatically applied to meet DATA Act objectives.
American Council for Technology-Industry Advisory Council (ACT-IAC)

The American Council for Technology (ACT) is a non-profit educational organization established in 1979 to improve government through the efficient and innovative application of information technology. In 1989 ACT established the Industry Advisory Council (IAC) to bring industry and government executives together to collaborate on IT issues of interest to the government.

ACT-IAC is a unique, public-private partnership dedicated to helping government use technology to serve the public. The purposes of the organization are to communicate, educate, inform, and collaborate. ACT-IAC responds to government requests using a model that includes government and industry working together, elbow-to-elbow. ACT-IAC also works to promote the profession of public IT management. ACT-IAC offers a wide range of programs to accomplish these purposes.

ACT-IAC welcomes the participation of all public and private organizations committed to improving the delivery of public services through the effective and efficient use of IT. For membership and other information, visit the ACT-IAC website at www.actiac.org.

Collaboration & Transformation SIG Financial Management Committee

DATA Act – Transparency in Federal Financials Project

The C&T SIG sought input from the Department of the Treasury and the Office of Management & Budget (OMB) in following the progress of the Digital Accountability and Transparency Act of 2014 (DATA Act, P.L. 113-101) from the pilot phase through practical/production implementation. This report provides useful information for industry and government managers to consider as they assess their readiness and develop their strategies to meet the new requirements.

Disclaimer

This document has been prepared to provide information regarding a specific issue. This document does not – nor is it intended to – take a position on any specific course of action or proposal. This document does not – nor is it intended to – endorse or recommend any specific technology, product or vendor. The views expressed in this document do not necessarily represent the official views of the individuals and organizations that participated in its development. Every effort has been made to present accurate and reliable information in this report. However, ACT-IAC assumes no responsibility for consequences resulting from the use of the information herein.

Copyright

©American Council for Technology, 2015. This document may not be quoted, reproduced and/or distributed unless credit is given to the American Council for Technology-Industry Advisory Council.

Further Information

For further information, contact the American Council for Technology-Industry Advisory Council at (703) 208-4800 or www.actiac.org.
Introduction

On May 9, 2014, President Barack Obama signed the Digital Accountability and Transparency Act (DATA Act, P.L. 113-101). With this mandate, the Department of the Treasury (Treasury) and the Office of Management and Budget (OMB) are required to transform U.S. federal spending information sharing and reporting from non-standard data and non-integrated documents into open, standardized, machine-readable data – accessible online to the public, via an improved USASpending.gov portal. “Federal spending” includes the entire spending lifecycle in detail, from appropriation through assignment and disbursement of grants, contacts, and other administrative spending.

In short, the DATA Act is both a mandate and a challenge for all recipients and reporters of government spending to clean the data, align information-sharing systems and structures, and generally improve the IT context that will enable compliance. This is not a new kind of challenge, however – existing federal legislation regarding information sharing and usefulness already demand more scrutiny and quality improvement around the data that’s used and is to be more transparent – from the Federal Data Center Consolidation Act of 2013 and the Government Performance and Results Act of 2010 to the more recent Digital Government Strategy and Open Data Policy.

What’s different, however, is rapid and broad exposure to many internal and external stakeholders of the following IT infrastructure investment priorities that collectively represent significant keys to IT simplification:

- **Consolidation**: the DATA Act mandate includes no new budget, yet requires agencies to instigate or take advantage of a wealth of shared services and data management improvement or modernization programs already underway, both in government and industry, to reduce duplication and unnecessary IT management and integration complexities
  - Message: “standardization and consolidation initiatives are a priority, aligned via enterprise architecture tenets.”

- **Engaged governance**: most public sector agencies are faced with generational data management change drivers already, from big data to secure mobile analytics requirements. This federal-led initiative provides a top-down, organizational imperative for actionable, cost-effective data governance across the entire community of data users and stewards
  - Message: “let’s get committed, transparent, and hands-on with data governance.”

- **Virtualization**: the variety of data standards and processing maturity across all the stakeholders is so great, from federal to state, local, and private recipients, that the elements of a solution will require a great deal of abstraction from the legacy data stores, systems and acquisition plans that can’t easily be changed. This introduces a dynamic, agile layer of usefulness between the existing IT infrastructure and new users with high, consumer-driven expectations.
  - Message: “do no harm, but expose tangible value quickly.”
IT simplification is a theme that ties together these investment priorities – but it isn’t really a new idea – more so an idea that’s not only mandated but is absolutely achievable in this era of cloud-enabled computing platforms. A typical roadmap to IT simplification begins with *standardization* of components, evolves to *consolidation*, then to *optimization*, leveraging virtualization techniques, and ultimately to *utility computing* (IT as-a-service, sourced within or through a cloud provider). An [IDC White Paper Simplifying IT to Drive Better Business Outcomes and Improved ROI: Introducing the IT Complexity Index](#) illustrates how this roadmap works, identifying “the need to take an entrepreneurial approach – replace outdated infrastructure or building out new IT to support new business initiatives…virtualization can help pilot this, test it out, migrate, perform more ‘agile’ software development and rollouts…being able to respond and collaborate more quickly with users.”

Note that many of an organization’s existing IT Infrastructure systems, capabilities, and programs are likely to be involved or impacted in some manner by such an end-to-end data management challenge. Per the [DATA Act Playbook](#), this data-centric approach “differs from the traditional system-centric way of collecting, aggregating, and validating additional data from agencies via a bulk file or aggregating information in a central system, never to be reused by the agency.” Such a data-centric approach will not only drive new capabilities and services on the external consumption side, but also on the agency-internal consumption and production side.

The DATA ACT Playbook itself alludes to planning for system and business process changes, along with implementation of the “broker” concept – an abstracted or virtual data layer. Some systems may be new, or be retired; some may not be able to be changed at all. Any system changes, however, may be facilitated or even obviated, both in terms of IT and business or mission performance, via focused attention to the enterprise-wide IT infrastructure investment priorities described above (i.e., consolidation, applied data governance, and virtualization). In fact, implementation of playbook steps 4-7 – *Design Changes*, *Execute Broker*, *Test Broker*, and *Update Systems*, will all benefit from consideration of these priorities.

So who’s involved in deciding and driving consideration of these IT investment priorities? Who are the change agents for IT simplification, and therefore the primary audience for this paper? The hands-on stakeholders - those whose roles, skills, and responsibilities may change the most given the DATA Act - of an organization’s IT Infrastructure capabilities usually operate within three contexts:

A. **IT architecture** (planning and modeling),
B. **IT governance and compliance** (standards, agreements, and authorities, including security), and
C. **IT engineering and operations** (design, build, and run).

**Recommendations**

1. **IT Infrastructure Consolidation**

From initial receipt, ingest, or entry of data relating to the receipt and use of government funds, a grantee organization relies on multiple systems, technologies, and interfaces to manage the...
The data includes not only spending amount and status values, grouped and
categorized by a financial taxonomy (e.g., transactional data), but also attributes/metadata
referencing the data itself, such as the data’s age, security, provenance, etc. The data is either
“source” or “system of record” data, or something else – like derivations, transformations,
visualizations, referenced authoritative values (e.g., reference data), etc.

Different approaches by organizations in the storage, management, production, and
consumption of the data have resulted in a wide variety of technology infrastructure elements
necessary to support – from very complex, granular and highly-customized infrastructures to
simple, standard packagedollections of hardware and software assets. The proliferation of
individually-configurable IT elements and services often makes it more difficult for an
organization to be agile, unable to reduce costs while maintaining performance or to respond
efficiently to opportunities or threats.

This includes responding to highly-publicized challenges the DATA Act represents, meeting the
expectations of multiple and diverse stakeholder groups – essentially an entire “open data”
community. Challenges relating to siloed or fragmented data, duplicative data processing
investments, legacy IT with a very high “total cost of ownership” (TCO) – yet with very opaque or
simply unknown effectiveness (return on investment or “ROI”). It’s simply too difficult or
expensive in many IT environments to consistently track and record the entire lifecycle of a data
element across multiple systems, repositories and interfaces, if this wasn’t the design in the first
place.

How is IT consolidation to be approached, to benefit the organization overall, while
directly facilitating the DATA Act objectives?

Identifying the primary targets for IT infrastructure consolidation, and aligning the target priorities
to existing systems upgrades, technology refreshes, or development underway should reveal
very specific consolidation opportunities with both long and short-term ROI. Identifying these
targets is an enterprise or solution architect role – for someone or a group that knows or
understands the inter-dependent relationships and status of the various IT systems and services
and their acquisition status. Some of these targets will be unique to the organization while
others will likely include some of the following areas of opportunity that reflect common,
repeatable needs, such as listed below.

**IT Consolidation Targets of Opportunity in Order of Highest Anticipated ROI**

- **Data management repositories and processes** (e.g., input/update, move, store/distribute,
archive processes that populate the data repositories, etc.) – in simple terms, this is the
consolidation of databases and file/content management systems, which is fast becoming a
critical priority for most IT departments, given tremendous growth curves in collected and
ingested data. Consolidation of data repositories to address big data challenges in particular
is driving awareness of “data reservoir” concepts, which highlight new methods and tools for
cross-repository data quality change, management, and reporting requirements, such as “big
data SQL” (standard SQL access to all consolidated datastores, enabling bridging between
RDBMS, Data Warehouse, Hadoop, and NoSQL implementations).
• **User visualizations** – the proliferation of portals or consoles for reporting or viewing data and information can put significant risk into maintaining the timeliness and veracity of data across these points of interaction, as well as introducing unintended, excessive consumption on the underlying computing infrastructure, when user query resource consumption is replicated and needlessly duplicated, vs. leveraging shared resources.

• **Software platforms** (e.g., database, middleware, etc.) – centralizing the management of data and middleware environments will reduce the amount of resources needed to integrate and continue to maintain disparate technologies, and free IT staff to focus on higher-value work.

• **System interfaces** (e.g., all interface types, from fixed to service-oriented) – the consolidation at the middleware layer, seeking any opportunity to reuse web services and standard interface schemata, simplifying and consolidating service directories and registries, as well as consolidating versions of integration software to manage and monitor.

• **Secure auditing/logging/analytics** – the consolidation of secure audit logs, system and application performance, user and interface activity, information lifecycle events, and data analytics should enable more rapid and efficient introduction of change, and analysis of the response to it. In fact, in a recent interview Joe Hungate, Deputy CFO, Department of Housing and Urban Development, described a projected reliance on a consolidated data analytics platform to “tie everything together.”

• **Data quality** – the consolidation of data quality infrastructure that supports a data quality lifecycle, from auditing and profiling, to standardization, matching, merging, and verification, in addition to enabling access both as an asynchronous and real-time service, will help not only maximize investment in this tooling, but also focus and centralize the knowledge of these tools and their use in a “center of excellence” context. By “data quality infrastructure,” we mean the software and hardware used to plan, configure, design, test, and execute data quality lifecycle processes. This may include interfaces to multiple authoritative sources, temporary cleansing and staging repositories, reference and master data repositories, and reporting consoles. Note that data quality infrastructure may ultimately be leveraged across multiple data management domains, improving the signal-to-noise ratio everywhere, including:
  
  o Data repositories; permanent, transient, or temporary, for both “core” and metadata,
  o Queries/reports/searches, both input and output, in terms of user interfaces, and
  o Audits/logs.

• **System functions, services, and automated business processes** – services-oriented architecture (SOA) elements that may inherently be reusable and shared, but are not presently, and where multiple groups of users exist with significantly overlapping roles, responsibilities, and data stewardship functions. This speaks to some degree of SOA governance maturity, to both discover and manage reusable IT service components across and between organizations, and progress made here can directly facilitate the change required to address DATA Act requirements.
The DATA Act – IT Infrastructure Guidance

- **Data movement** – facilitation of “agile data movement” (on demand and in real-time), leveraging a consolidated data movement infrastructure across both temporary (test, cache, disposable, perishable, and data reservoirs) and persisted repositories (permanent storage tiers, archives, warehouses, and backups) can significantly reduce the risk of lost or misused data, particularly among engineering environments, as implemented for release management and testing cycles.

- **Consumer/government off-the-shelf (COTS/GOTS) applications** – consolidating applications, for example financial systems in particular, may in fact be both a driver and a facilitated outcome of other IT infrastructure consolidation plays, as described in the ACT-IAC *Transparency Enabling Transformation white paper*: “For those agencies who have not yet begun, the DATA Act can serve as the impetus for a strategy for interoperability in financial reporting. Effective implementation of the DATA Act will drive the consolidation and interoperability of operational and program systems, leading to substantial cost savings for agencies. Data consolidation and the interoperability of systems leads to more effective and efficient systems that are easier and less expensive to maintain and support.”

- **Data storage** – storage consolidation, as well as the consolidation of associated data management and access systems or services, should result in fewer opportunities for loss of data quality, and less effort and risk in expanding or updating data quality initiatives.

- **Compute** – consolidation and standardization of compute platforms, whether consolidation of physical assets or virtual instances, can directly reduce the cost and complexity of system updates and change, including testing data processing software and services.

2. **Engaged Data Governance with an IT Governance Perspective**

This mandate is a catalyst for enterprise data governance, though to be implemented or accelerated in a manner that delivers rapid value and compliance, while setting the foundation for persistent maturation and improvement. Such agility applied to an organization-wide imperative therefore requires a multi-disciplinary audience and governance framework - across the entire lifecycle of a data asset within and beyond the organization.

A fully-engaged (useful to and used by ALL stakeholders, within and outside of your organization) data governance framework is recommended to have four spheres of influence, or “swim lanes” of activity within a typical IT environment. Deploying a data governance framework in this way engages all business and IT communities associated with the changes triggered by the DATA Act, from new users and analysts to existing engineers and data stewards. The more stakeholders across the data management and governance ecosystem that are actively engaged in stewarding DATA Act changes through their respective IT infrastructure environments, the better.

There is, however, an implied organizational requirement that a formal steward of the enterprise data governance capability does exist, whether in person or in process (among contributing roles). This role is sometimes titled as “chief data officer” in organizations with a mature data governance process, or more likely a “chief data architect/steward” within the CIO organization. We recommend establishment of such a role, across the entire data governance ecosystem of DATA Act constituents.
Data Governance Spheres of Influence that Require Enterprise-wide Engagement

- **Sandbox**: data governance as applied to pilot/prototyping/sandbox activities, i.e., more relaxed, though with more immediate and collaborative feedback mechanisms and possibly public/private digital engagement methods, such as an open data or crowdsourcing portal. The application, monitoring, and support of this kind of governance is very flexible, with a community-managed mindset. Digital engagement methods useful for publishing solicitations for volunteer feedback, code or design reviews, data standards suggestions or examples, and possibly APIs to standards compliance automated testing services include agency social media channels, social-enablement of public portals or discussion groups, and public social enablement of document repositories. **Examples** of these kind of engagement methods and the supporting IT infrastructure are the [Java Community Process](https://www.jcp.org), used to test compliance with new Java standards, and the [OASIS Content Assembly Mechanism (CAM) Technical Committee](https://www.oasis-open.org/committees/cam), used to test compliance with NIEM standards.

- **Dev/test**: data governance as applied to system development, IT procurement, or technology refresh initiatives already underway – low-impact, non-intrusive, iterative improvement, seeking any opportunity for smart acceleration. Recommendations and decisions about data standards, processes, or architecture design are more guidance than prescriptive, seeking to mitigate future or likely risk. This kind of governance from the DATA Act perspective may involve addition of a data steward/architect focused on DATA Act requirements into the standard engineering review process.

- **Targeted**: data governance as applied specifically to new initiatives concerning specific DATA Act-identified data entities, for example, a focused cleanup, introduction of a new attribute or field, or creation of a new dashboard or mashup application. Here is where the existing engineering review process, sometimes instantiated as an architecture review board or data quality review, should be updated to specifically address and include DATA Act standards and compliance.

- **Enterprise**: data governance as applied at the enterprise level, across ALL data entities, enabling organizational data governance improvement. In order to most effectively support the changes the DATA Act requires, while taking advantage of the awareness and focused interest in data governance as a discipline generated by the legislation, tangible data governance tools and services may be a wise investment at this time, or if the investment’s already been made but is underutilized, a wise re-investment of time and energy to deal with this change. Data governance tools are as important to implementing new systems and services as they are for controlling risks in large data migrations, conversions, consolidations, and platform upgrade initiatives. These tools, however, need to be paired with a supporting data governance organization: a documented, accessible, and trained community of data practitioners and stewards from all areas, linked through a common, effective, and agile decision-making process regarding data standards and compliance.

Products providing data relationship governance, for example, can be utilized across the entire enterprise to provide the change management and data quality remediation workflows essential for users, analysts, subject matter experts, and signing authorities. These workflows and data management tools enable the collaborative creation, correction, and
conformation of master reference entities, attributes, relationships, and mappings so they're fit for purpose across multiple business objectives, including conformance to the DATA Act, yet are bound by community-wide referential integrity.

3. IT Virtualization

IT investments, planned or underway, will all require evaluation against a new, common set of enterprise data technical requirements or "enabling guidance." This guidance must align to the needs generated by new, mandated enterprise data standards and other stakeholders in the data governance community – both “open” and dependency requirements – and align with the entire data management lifecycle.

Evaluation of technical requirements will result in technology allocation decisions – IT products, solutions, or services that need to be procured, updated or upgraded, reused or repurposed, extended or enhanced, etc. Simply put, IT change is the outcome of the DATA Act’s actual or inferred technical requirements.

How can organizations mitigate negative impacts to existing production environments and processes while introducing these DATA Act-driven IT changes?

One method may be to determine whether virtualized technology resources can be used for each kind of technology requirement that will result in a technology allocation decision, and therefore whether a "cloud" deployment model for these resources is warranted.

IT virtualization is essentially an abstraction layer between functional compute logic and user interfaces, and the IT infrastructure – or components thereof – required for management and delivery. Enabling users to self-provision virtualized hardware and software resources for their compute workloads is core to the definition of cloud computing. This deployment model for virtualized resources may in fact be a business requirement of an organization that influences, or is driven by, the organization’s IT virtualization priorities.

Technical requirements may include virtualization of segments of the IT infrastructure, but may or may not result in a cloud capability – i.e. an on-demand technical service delivered via internet protocol to public or private, internal consumers. IT virtualization provides many scaling and rapid deployment benefits that may facilitate the core IT infrastructure change needed to accommodate DATA Act system, data, functional, and process changes. Virtualization opportunities across an enterprise IT infrastructure environment can be introduced in many areas, likely many more than your organization is considering today, and have been generally organized as follows:

- Compute platform virtualization (hardware platform)
  - Data storage
  - Operating system (OS)
  - Compute hardware (server)
  - Networking (software-defined)
  - Mobile device
• Application software virtualization (software platform)
  o Middleware/services
  o Database
  o File system
  o Applications
  o Desktop
• Data virtualization

Each of these areas of IT virtualization may facilitate an organization’s response to DATA Act technical requirements; we’ll explore data virtualization in particular.

Data Virtualization

Open data initiatives have encouraged innovation for improving government accountability and transparency. Government data portals are now evolving into information/knowledge hubs that organize and curate the data to support decision-making. There are many stakeholders who benefit from open data, including the government. The nature of innovation is that developments often come from unlikely places. The objective is to provide the access to government-held data to the public and entrepreneurs while appropriately safeguarding sensitive information and protecting privacy. Some providers of broadly-used technology and data services (i.e. the major Internet search engines and social media communities) showcase ease of access to their published APIs, resulting in a large ecosystem of partners and business opportunists. The government demonstrates similar evolving best practices through initiatives like USAspending.gov and Data.gov.

Data virtualization, however, is a somewhat broad and difficult term to define, though general consensus deems it to include “any approach to data management that allows an application to retrieve and manipulate data without requiring technical details about the data, such as how it is formatted or where it is physically located.”1 This concept and approach is critical for organizations with very large and old repositories of siloed data that will require examination for change requirements pursuant to the DATA Act.

For example, the Department of Veterans Affairs Office of Technology Strategies is developing Enterprise Design Patterns including a hybrid data access pattern currently termed “Data-as-a-Service”. This enterprise architecture guidance is focused on the data virtualization concept and possible solutions, which would include best practices for abstracting useful, accurate, and current data from underlying separate and heterogeneous technologies (i.e., “change data capture”), in order to ensure enterprise-wide availability of authoritative data sources and service member records.

1 https://en.wikipedia.org/wiki/Data_virtualization
Applying DATA Act standards requirements at this abstraction layer, made possible by data virtualization techniques, is a viable solution to the challenge of rapidly testing and applying standards changes to older systems.

To be more specific, delivering DATA Act-instigated data virtualization may include:

- Providing metadata or standards via APIs,
- Providing API configuration controls and maintaining version compatibilities,
- Devising a mechanism to disseminate data standard/API changes through data feeds,
- Providing examples of useful and successful APIs, and
- Providing toolsets, artifacts, and libraries that support data translation from common and prevalent data formats to Treasury adopted data standards.

This helps the agencies that need to report data as well as external stakeholders that consume the published data. To ensure such an "open data as-a-service" objective is successful, efforts to foster the exposure and use of reusable data from federal data repositories like USASpending.gov are critical. Examples of these kinds of efforts and the resulting capabilities include:

- **Mash-ups**: this is the process of taking the data from the federal sites and blending into data from other sites (private or other public). One such site that has used government data is WashingtonWatch.com, which combines government calculations about the costs or savings from proposed changes to government spending, taxation, and regulation and combines this data with the website’s own calculation of the net present value. These types of mash-ups allow the data from one system to be compared to another. Other examples include OpenCongress.org and AppsforDemocracy.org. These types of sites ideally operate through direct APIs to the source data or through JavaScript Object Notation (JSON) or similarly standardized, accessible digital files provided to the developer. An example would include the open data catalog data sets published by DC.gov.

- **Data analysis**: the use case for data analysis is the process of an organization downloading the data and then loading the data into analysis tools. In most cases, the data is loaded into some type of database, which is then analyzed using the analysis tools, which can be very complex or fairly simple.

- **Dynamic linking to authoritative data**: in some cases, a system would like to interchange with the federal government to have access to the same authoritative source data or information. For example, a business may require access to the universal award ID to determine if the award IDs entered in the systems are valid. Protocols such as RDF or ODATA on top of a REST API would provide a standardized, effective method for users to effectively use the government reference data. Authoritative data may include not only the discrete values and metadata, but also the data structure containers and standards – such as the XML schemas managed by the NIEM program.
Discoverable Information Catalog

A data-centric vision for the implementation of the DATA Act results in a discoverable information catalog that is centrally published on a repository such as USASpending.gov. The catalog will then be used by several different stakeholders (federal agencies, state, local, citizens, and industry) in various access mechanisms to search and discover all information published as part of the DATA Act. A search engine around spending information would be widely beneficial. Search technologies include not only user queries and faceted navigation (exploration by topic) of data repositories, but also real-time monitoring and pattern or correlation notifications that users can configure and subscribe to. An obligation of funds to a particular entity, for example, correlated with the timing of related material procurement funding at a different entity, could generate a data event that might be useful for the industry in tracking resource availability and consumption.

IT Virtualization Use Case

Following is a use case of how an IT Virtualization investment perspective supports a DATA-Act induced change. Consider these general business requirements or expectations concerning the DATA Act, for federal agency financial management and payment systems:

- Agency alignment with the new proposed DATA Act expanded data schema as defined by Treasury and OMB (beyond FFATA requirements).
- Ensuring current federal agency financial management and payment systems are upgraded to capture the fidelity of data requirements as required by the DATA Act.
- Ability for federal agency financial management and payment systems to provide extracts of the DATA Act required data on detailed financial transactions in the desired format on a prescribed frequency basis.
- Define and enable interfaces between federal agency financial management and payment systems to export or provide real-time data feeds to USASpending.gov systems in the prescribed DATA Act schema format and desired frequency.
- Ability for federal agency financial management and payment systems to interface with repositories and systems that are the primary sources of the DATA Act objects, entities, and transactions to ensure consistency in defining and reporting on unique entities across all agencies.
- Ability for federal agency financial management and payment systems to be governed by a data governance board for the DATA Act, and to be flexible to adapt and iterate on the DATA Act schema and functional requirements on an ongoing basis.
Taking the chief data engineer/architect’s perspective, for example, the nature of these business and functional requirements would likely result in or influence technical requirements in the following main areas of concern, among others:

- **Data management** (run-time/ops)
  - Requirement: the new data standards must be maintained and validated across the entire data management lifecycle (within and between these financial management and payment systems), from the point the data is sourced or introduced to its ultimate disposition.

- **Data engineering** (dev/test/prototype)
  - Requirement: the new data standards must be able to be maintained and validated in each IT engineering environment that is used to update these systems, as the system changes are planned, designed, tested, and deployed.

These technical requirements are further validated and detailed by the IT engineering and operations leadership, or "infrastructure stewards" with whom the data engineer collaborates, who are most knowledgeable about the inventory and status of the enterprise’s IT infrastructure assets.

How can these technical requirements be delivered or influenced through the introduction of IT infrastructure virtualization mechanisms?

**Data Management Technical Requirements Facilitated through IT Virtualization**

The solution, whether COTS, GOTS, or updates to an existing system, must protect and preserve compliance with the data standards through the entire data management lifecycle for the citizen’s data: from initial interface with the data source (ingest/interface), through intermediary ETL (extract, transform, and load) processes, within managed data stores (whether NoSQL, RDBMS, file systems, Warehouses, etc.), through any information/data service interfaces, and ultimately as presented or served via user or endpoint machine interfaces.

Virtualization of data sources (data as-a-service or data virtualization), at any stage in the data management lifecycle, can rapidly expose the data for analysis, correction, reporting, visualization, or further processing, in compliance with the data standards – without necessarily impacting or changing the source systems. Data virtualization techniques are an extremely effective way to introduce data quality methods and tooling into an existing IT infrastructure as well – to profile, correct, and further share or analyze the data without necessarily impacting the existing source systems. In essence, to perform or test data management processes on replicated or virtualized data.

This IT infrastructure virtualization technique can therefore facilitate rapid, risk-mitigated delivery of the data management technical requirements that the DATA Act business requirements will generate.
Data Engineering Technical Requirements Facilitated through IT Virtualization

New data models, whether for persisted data or metadata, data-in-transit (via system interfaces, including messages, file transfers, and APIs), or for user interfaces (visualization, data entry, or reporting) must conform to the new DATA Act standards both in execution and in testing/validation; there must be a common method and infrastructure to test or prototype data standards compliance during system development, as well as evaluation and validation of data standards compliance when systems execute.

This common testing method and infrastructure might not be developed internally, but may be found via utilization of exposed shared services from another agency or community stakeholder, and this means new external interfaces to be built and configured within the dev/test environment, which can be a challenging networking and security requirement for IT departments.

Virtualization of compute and application platforms to quickly create and discard test or prototype environments including servers, databases and middleware – possibly without the overhead of standard configuration management compliance – can provide a mechanism for the rapid validation of innovative approaches, or of changes targeted for sensitive areas of existing systems. This kind of cloud-deployed virtualization capability may in fact be very helpful to rapidly create temporary environments to plan and test major system changes or data migration methods, or rapidly create more permanent environments for recurring data cleansing, monitoring, and analytic tasks.

This IT infrastructure virtualization technique can therefore facilitate rapid, risk-mitigated delivery of the data engineering technical requirements that the DATA Act business requirements will generate.

Authors & Affiliations

Herschel Chandler, Information Unlimited, Inc.
Subhasis Datta, ICF International
Ted McLaughlan, Oracle
Sumit Shah, CGI Federal