

RESEARCH

Open Access



A comprehensive model for management and validation of federal big data analytical systems

Feras A. Batarseh^{1*}, Ruixin Yang¹ and Lin Deng²

* Correspondence:

fbatarse@gmu.edu

¹College of Science, George Mason University (GMU), Fairfax, VA, USA
Full list of author information is available at the end of the article

Abstract

Background: In this era of data science, many software vendors are rushing towards providing better solutions for data management, analytics, validation and security. The government, being one of the most important customers, is riding the wave of data and business intelligence. However, federal agencies have certain requirements and bureaucracies for data-related processes, certain rules and specific regulations that would entail special models for building and managing data analytical systems. In this paper, and based on work done at the US government, a model for data management and validation is introduced: Federal Model for Data Management and Validation (FedDMV). FedDMV is 4-step model that has a set of best practices, databases, software tools and analytics. Automated procedures are used to develop the system and maintain it, and association rules are used for improving its quality.

Results: After working with multiple engineers and analysts at the federal agency, there is a general consent that FedDMV is easy to follow (please refer to the experimental survey). However, to quantify that satisfaction, three experimental studies were performed. One is a comparison to other state-of-the-art development models at the government, the second one is a survey that was collected at the government to quantify the level of satisfaction regarding FedDMV and its tool; and finally, a data validation study was performed through detailed testing of the federal system (using an Association Rules algorithm).

Conclusions: To develop a safe and sound federal data analytical system, a tested and rigorous model is required. There is a lack of government-specific models in industry and research. FedDMV aims to provide solutions and guided steps to facilitate the development of data analytics systems given the governmental constraints. FedDMV deals with unstructured data that streams from multiple sources, automates steps that are usually manual, validates the data and maximizes its security. The results of the experimental work are recorded and reported in this manuscript.

Keywords: Data management, Big data analytics, Validation, Unstructured data, Federal agency

Background

The promise of intelligent and accurate predictions in software systems that was previously pursued by many [1, 2] is now being transformed into a new testament of big data analytics. Undoubtedly, data analytics is a buzz word, if not the “*buzziest*” word of the day [3–6]; it is a science that lies in the nexus of Statistics and Artificial Intelligence (AI). The

goal of data analytics is to return an intelligent and more focused version of large and impersonal datasets, provide quick insights into data, and help with visualization and decision making. Literature has many examples of successful applications of data analytics, not only to specific business-driven functions, but also many industrial and research domains such as Healthcare [3], Education [4], Banking & Finance [5], and many others. The Government is no different. Actually, most federal agencies understand that data can help them unlock the government of the future towards better operations, better citizen service, and more efficient decision making.

Data engineering at the US government

In 2012, the Big Data Initiative [7] (refer to the screenshot shown in Fig. 1 below) was created; that was driven by an earlier Open Government initiative signed in 2009 [8–10]. These two initiatives aimed to create data openness and ensure machine-readable data at all federal agencies. Accordingly, multiple agencies started sharing their data through outlets such as www.data.gov. The Open Data Initiative announcement's was reinforced in 2013 as an executive order.

The two initiatives created a new reality within federal agencies: an instant need to implement a data system that would allow for managing, browsing, mining, and sharing their data. In this manuscript, a data system is defined as:

1. A group of databases
2. Tools that manage the databases
3. A set of validation, security and safety procedures
4. The outcomes of the data system such as dashboards and visualizations.

Based on this novel direction, federal agencies are racing to implement data analytical systems, this paper aims to guide these agencies through the process – using the proposed federal-specific model. Although the private sector is known to be the main



Fig. 1 The Big Data Initiative Announcement's Header [7]

driver of technological advancements, the increasing trend of software engineering in the government is undeniable [11], and is driving the change as well. Previously, software development at the government had followed the generic waterfall model [12], where everything is sequential. That incurred massive costs and time delays. Based on an article by the Washington Post [13], a monetary value of up to \$50 million was wasted on failing federal software projects (in one year). Many governmental agencies (such as: the Financial Industry Regulatory Authority and the Department of Health and Human Services, and many others) still latch onto the expensive and difficult to manage waterfall model (the Affordable Care Act website www.healthcare.gov is a recent example [14]). Overall, federal agencies spend around \$77 billion dollars on software and IT every year [15]. Based on that, the United States Government Accountability Office (GAO) published an important report [15] making the case for the need to switch government software engineering practices to agile [16]. More importantly, the GAO report identified 14 challenges that governmental agencies undergo. Shown in (Table 1) below, the challenges are specific to governmental workflows; most challenges are associated with Collaboration, Teamwork, and old fashioned federal practices. This fast movement towards agile software development and data analytics is the major motivation for our work.

In a thesis/study published at Princeton University, data structuring was identified as one of the major challenges at federal agencies [17, 18]: “When government does collect and publish data in a reusable way, government enables third-party stakeholders like advocates, academics, journalists and others to powerfully adapt its data in any way they see fit using the latest technologies, and to add value in unexpected ways. Third parties can use government data to experiment in parallel, in order to discover what innovations work best in changing technological environments”. To engineer better data systems however, it is important to first understand the nature of data available in the government; the next section discusses existing and previous federal experiences in this realm.

This paper is structured as follows: the next sub-section looks into what has been already done in the government in terms of software engineering and big data analytics. The section after discusses related work and current federal challenges. Afterwards, the main contribution of this paper, the Federal Data Management and Validation Model (FedDMV) is presented. Subsequently, two experimental studies and survey (on the usability of the tool) are introduced; and the last section concludes the paper with results and future work.

Table 1 GAO’s Fourteen Agile Federal Challenges

Table: Federal Challenges	
Teams had difficulty collaborating closely.	Procurement practices may not support Agile projects.
Teams had difficulty transitioning to self-directed work	Customer did not trust iterative solutions.
Staff had difficulty committing to more timely and frequent input.	Teams had difficulty managing iterative requirements.
Agencies had trouble committing staff.	Compliance reviews were difficult to execute within an iteration time frame.
Timely adoption of new tools was difficult.	Federal reporting practices do not align with Agile.
Technical environments were difficult to establish and maintain.	Traditional artifact reviews do not align with Agile.
Agile guidance was not clear	Traditional status tracking does not align with Agile.

Unstructured data at the US government

To address a legislative issue, for decision making, and for media announcements, most federal agencies require data that streams from other federal departments; or in some cases even city, county or state governments. Some of that data is shared with the public.

Besides, the aforementioned www.data.gov, the PACER system (public courts online access system) [17], the BLS (Bureau of Labor Statistics), and NASS (National Agricultural Statistics Service) [19] are major examples of federal data system that adopted the open data initiative and made their data public to citizens. States and cities started publishing data to the public as well. For example, the cities of Chicago [20] and San Francisco [21] both have open data portals, data.sfgov.org and data.cityofchicago.org respectively. However, different parties share and store the data in different formats, varying standards, and using different technologies. How is federal data used collectively then? How is it shared across agencies and governments?

Due to the size, unstructured and non-repetitive nature of federal data, Hadoop could be very helpful; Swish Data (among others such as Map Reduce and Horton Works) [22–24] published an article identifying multiple use cases for the government using Hadoop. Besides the ability to organize enormous amounts of size of data, Hadoop also handles unstructured data (which is a major characteristic of federal data). The government is indeed focused on using commercial data management tools such as: Microsoft’s SQL servers [25], SAS [26], and Tableau [27] (among others). Nonetheless, and regardless of the tool of choice, following a controlled data management model reduces the risk of building faulty data systems; it also ensures the ability to learn from other experiences and not repeat the mistakes of the past [28] (among many other advantages). It is apparent that big data is key for the US government (Table 1 and Fig. 2 illustrate use cases that show why big data is an inevitable reality at the government), however, what tools is the government using to manage its data? Besides relational databases, SQL, Oracle tools and the traditional Database Management Systems (DBMS), is the government using the *de-facto* of big data – Hadoop? Do they have specific tools that can manage and adhere to federal rules? No strong evidence of that is reported.

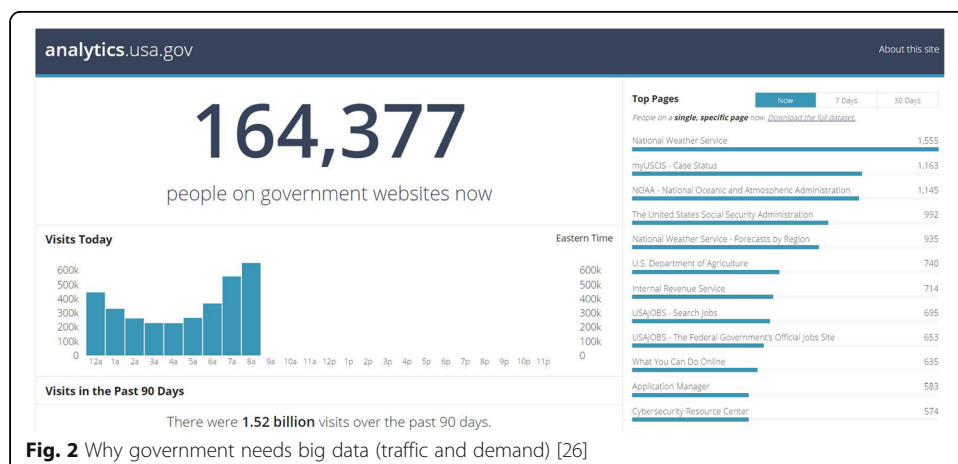


Fig. 2 Why government needs big data (traffic and demand) [26]

Related work: data management models

Lifecycles in software engineering have lead major shifts in the tech progress of the engineering world. The transition from Waterfall development to Spiral [12], from Traditional to Agile [16] and from product-based to contextual user-based [29, 30] development left an obvious fingerprint on the world of engineering. Do these models apply to developing a data analytics system? Although considered a fairly novel field, this section introduces the most prominent models for data analytical systems and how they are utilized.

Data analytics development for non-federal systems

Before federal got on the analytical wave, many industries (such as: healthcare, finance, athletics, and the media) adopted analytical models within their organizations. Although data mining research has been of interest to many researchers around the world, data analytics didn't see much light until it was adopted by the industry. Many software vendors (SAS, SPSS, Tableau, MSTR, Qlik, and Pentaho) shifted the focus of their software development to include a form of data analytics, big data, data mining, statistical modeling and data visualization. Based on multiple long and challenging deployments in many fields, trials and errors, and multiple consulting exchanges with many customers from many fields, these vendors coined a data management model for data analytics. SAS (based on Gartner's research [31] is one of the pioneer vendors in this field), provided that model (illustrated in Fig. 3). The model includes the following steps:

- a. Identify and formulate the problem
- b. Prepare the data (pivoting and data cleansing)
- c. Data exploration (summary statistics, bar charts and other means of exploration)
- d. Data transformation and selection (select ranges, and subsets)
- e. Statistical model development,
- f. Validation and deployment
- g. Evaluate and monitor results of model with data
- h. Deliver and refine the model.

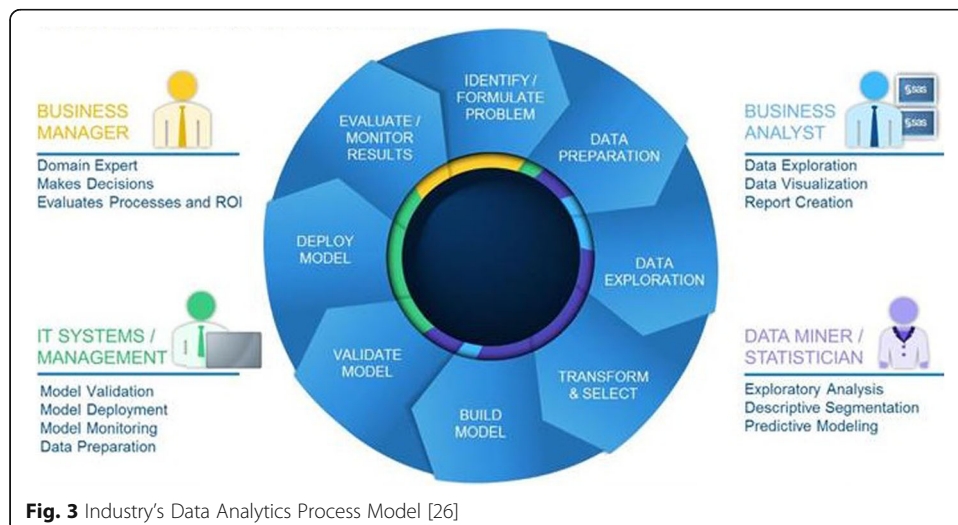


Fig. 3 Industry's Data Analytics Process Model [26]

When software vendors like Tableau, SAS, SPSS and others realized that the government is highly interested in data analytics as well, they started to propose different models to the government to win its projects. Next section discusses some of these proposed models, and later sections of the paper introduce our new model: The Federal Data Management and Validation Model (FedDMV), and contrast it with an existing relevant model.

Federal data management, validation and analytics

As it is already established, federal departments and agencies are very good with generating and collecting data, but not as good in storing and sharing it. Government IT however, is being constantly challenged to make this data available to employees, the media, the public and other agencies. To address such challenges, multiple software vendors produced federal-specific solutions. Actuate [32] for example introduced BIRT (a data driven model). It is one of the most used models in government. It has been used by more than million users and features an active community, as Actuate claims [32]. The BIRT process matches most federal regulations and certifications. BIRT provides a list of reports and dashboards that agencies can plug into their data; it has dashboards that deal with federal data sharing, data fraud detection, performance management, and citizen self-service. Actuate deploys BIRT to support federal data operations at the Department of Defense, Federal Aviation Administration, US Equal Employment Opportunity Commission, and others. BIRT however, just like most “commercial solutions” is not a solution that is easy to manage without the support of Actuate, and it implies high dependence on the vendor. It has the advantage of providing dashboards, and quick-ready solutions, but the dependency limits the federal agencies and departments from having their own controlled data infrastructure. Another major player in the federal data area is Salient [33]. Salient calls their solution the Federal Mission Software Solution (FMSS). FMSS is based on agile development, it provides electronic workflows for data such as: electronic federal signatures, electronic disbursement of funds through the US treasury, electronic payment collection from citizens through www.pay.gov, and infrastructure management (among other workflows). Although some Salient solutions are deployed at the US Air Force, the US Army, Department of Commerce, and many others [33], Salient doesn't provide a comprehensive workflow to federal agencies that can be followed to build a data system, rather, the models by Salient are tailored to solve specific problems. The corresponding quasi-solutions to those problems are assembled, and accordingly Salient claims that the company provides a comprehensive federal solution for data analytics. Neither BIRT nor FMSS are fully sufficient to a federal agency or department to cover many data management requirements there.

Information builders on the other hand is a tool that provides a federal solution called WebFOCUSRStat [34]. RStat provides federal agencies with R-based advanced analytics to help with proactive decision making. Although multiple vendors provide advanced analytics through their solutions, RStat is specifically designed to federal agencies analytical needs. RStat focuses on the following fields for advanced analytics: Social Security, Tax Collection, Customs, Medicare/Medicaid, and Armed Forces. An example of a dashboard that federal users can interact with is presented in Fig. 4 below. RStat lacks descriptive and general types of analysis. It also lacks a sufficient process for data management that could be easily implemented by government workers.



Fig. 4 Federal Dashboard from the Department of Labor [34]

Other major players in the data analytics field for government are Deloitte, Accenture, SAS, Tableau and many others. Many of these vendors would be more than keen to get into the business of building solutions to the government, however, that means that they will be in control of maintaining, updating, and improving such systems, something that the government might not be able to afford. Due to a recent study by the Association of Government Accountants (AGA) [35], the government hasn't fully adapted data analytical infrastructures due to 3 major reasons: 1. A lack of budget resources (67%): getting any of the previously mentioned solutions or using any of the vendor-specific models exacerbates this problem. These vendor solutions are relatively expensive. 2. A lack of appropriate staff (53%) 3. Uncertainty as to how to develop a data analytics system (33%): deploying off the shelf solutions shall influence the agency in a negative way in terms of uncertainty [35]. Therefore, building an in-house solution creates less ambiguity and more clarity into the future of the data system at the agency. The AGA interviewed multiple representatives from multiple Federal agencies, they categorized the agencies into 3 categories based on desire, understanding and deployment of data analytical systems: High (23% of agencies), Medium (68%) and Low level (9%). Examples of the agencies and departments interviewed and are interested in a data management model for big data analytical systems: 1. US Department of Agriculture, Food and Nutrition Service (FNS): Deployed a data system called ALERT: a system for fraud detection. 2. US Postal Service: Deployed a system called RADR: a system that aids in selection of areas for investigation and audit. 3. Centers for Medicare and Medicaid Services: Deployed a system to identify and prevent waste, fraud and abuse in the Medicare fee-for-service program. Many other departments and agencies deployed data analytical systems, more are found under [35] and Fig. 5 below.

Therefore, there is a gap in literature for an in-house federal driven models that guide the implementation of a data system. Additionally, there is a lack of validation and data sharing techniques at the federal government.

ORGANIZATIONS INTERVIEWED	PURPOSE OF SYSTEM	STAGE OF DEVELOPMENT	EXTENT OF DEPLOYMENT
Recovery Accountability and Transparency Board Recovery Operations Center (ROC)	ROC – Identifies potential fraudulent transactions by analyzing ARRA funds information using 22 different data sets and analytics	Development Complete	Deployed across all ARRA funding, pilot testing against remaining federal funds
U.S. Department of Agriculture, Food and Nutrition Service Anti-Fraud Locator using EBT Retailer Transactions (ALERT)	Identifies food stamp trafficking by analyzing EBT transactions using pattern identification and risk rankings	Development Complete	Deployed and fully integrated into FNS operations
U.S. Department of Defense, Defense Finance and Accounting Service (DFAS) Business Activity Monitoring System (BAM)	BAM – Identifies improper payments before they are made by matching data across systems, looking for patterns of payments indicative of fraud from past transactions, etc.	Development Complete	Deployed across nearly all DFAS payment systems. Results are integrated into DFAS operations
U.S. Department of Defense, Defense Logistics Agency, Office of Operations Research and Resource Analysis (DORRA) Enterprise Business System (EBS)	EBS – Provides the DLA leadership with information on performance metrics based upon information from the agency’s systems	Development Complete	Deployed across DLA
U.S. Department of Defense, United States Navy, Naval Sea Systems Command, Office of Fraud Deterrence and Detection (NAVSEA-OFDD)	Planned system will be used to identify cases of potential fraud and abuse related to contract procurement	In Planning Stages	Not applicable
U.S. Department of Education, Office of Inspector General (ED-OIG) E-Fraud Analytical Model (EFAM) & State and Local Education Agencies Risk Model (SLRM) Systems	EFAM – Identifies fraud in higher education assistance programs by using data mining techniques SLRM – Analyzes information on federal funds provided to state and local education agencies to develop a risk model prototype	EFAM – System in Testing SLRM – Development Under Way	Deployment started.
U.S. Department of Health and Human Services, Centers for Medicare and Medicaid, Center for Program Integrity (CPI) Fraud Prevention System (FPS)	FPS – Analyzes Medicare payments, prior to payment, to identify potential fraud and abuse. FPS uses predictive modeling techniques.	Development Complete	Deployment across all Medicare fee-for-service payments; system integration has been ongoing for one year.
U.S. Postal Service, Office of Inspector General (USPS-OIG) Risk Assessment Data Repository (RADR)	RADR – Analyzes activity in four areas to identify potential fraud and abuse by using data mining and predictive analytics. Results referred to investigators.	Development Complete	Deployed and fully integrated into USPS-OIG operations

Fig. 5 Data Analytics Deployments at Federal Agencies and Departments [36]

That model needs to accommodate the following characteristics:

1. A model that facilitates data streaming between agencies.
2. A model that validates the data, maximizes its security and provides a standardization mechanism.
3. A model that is built in-house and without dependency on commercial products
4. A model that can handle big amounts of data

Based on what have been discussed in previous sections, and due to the immediate need for a management model that has the aforementioned four characteristics, this paper introduces: The Federal Data Management and Validation Model (FedDMV) to address all the challenges and fill the existing void. FedDMV has been used and experimented with at a federal agency. It is introduced in the next section.

Methods

The model for federal analytical systems

This section introduces the main components of FedDMV; the main contribution of this paper. The major overall challenge with big data is non-repetitive unstructured data. Structured data is much easier to manage, and if the data is repetitive, then it is also easier to predict its contents and therefore easier to control [36–38]. However, what is the best way to stream and structure data? When data resides at a federal agency, how is it validated? FedDMV provides answers to these problems.

The federal data management, streaming and validation

Besides guiding the development team through the steps of developing an analytical data system, FedDMV aims to solve three main issues:

1. Data volume challenges: Quantity of data that federal agencies import and generate is big (thus making it a big data challenge). This data is highly diverse and the speed at which data flows/updates from multiple sources is high (much of it is daily, some weekly, and some monthly/yearly).

2. Distribution challenges: The federal agency aims to build visuals and shareable solutions.

3. Data quality: Validating federal data has many challenges, and requires many resources.

FedDMV addresses these challenges and consists of the following main 4 steps (illustrated in Fig. 6):

1. Data Management:

- Data Sources Variables Exploration: all the unstructured data from all sources is unified into the new FedDMV centralized database.
- Variable Mapping and Columns Unification: redundant (repetitive) variables are concatenated in columns, and potential contents of the columns are identified (including their data types, size, and formats).
- Data Standards Creation: this step is the main step for creating Lookup (LU) tables in the database. LUs are dictionaries for standardizing the different forms of data from multiple sources.

2. Data Validation: Validate and verify the data in the target Database (DB) – using an Association Rules Algorithm (discussed in the next section). Validation is done in 3 steps:

- Test case collection and execution.
- Test case association measurement and evaluation.
- Data system refinement.

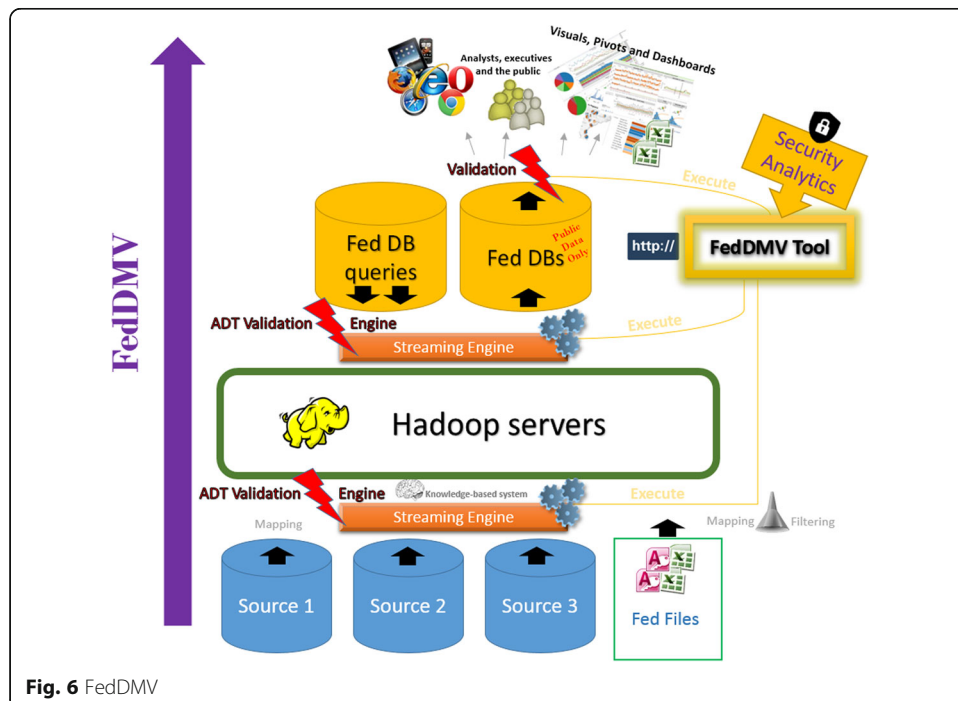


Fig. 6 FedDMV

3. Data Security Assurance
4. Federal Restrictions Deployment: deploy federal regulations into the data system.

All the steps are discussed next in great detail. Figure 6 illustrates the bottom-up workflow and outcomes for FedDMV. For the first step, and as the data flows in from different sources, it goes through a knowledge-base (KB). The KB explores the variables and organizes the data in a structured manner. The streaming process is shown in Fig. 7 below. The KB works as a mapping mechanism. Each variable from a source is mapped to a variable in the destination using a table of associated variables. For example, a certain commodity (auto part - tires) that is being exported from Germany to the US is represented with the following data string in one resource:

EXP TIRE, AUTO PARTS, GER, US \$70.00

Similar data is presented in a different form from another data source:

Auto Tires, Germany, Trade Exports from: USA, 70 US Dollars.

Another data source would have similar data like this: GRTIRExUS70; that needs to be broken down into multiple fields and eventually migrated into data fields in the destination table. The design for the database needs to consider both (and possibly more) formats of data streams. However, similar data fall within the same data group (the data is grouped into different types/groups and each group consists of data that share common characteristics). The goal is to standardize the data format, and to have a single version of the truth data format in the FedDMV system. The dimensions that need to be standardized cover the *where*, *when* and *what*. That is correspondent with *Geography*, *Time*, and *specific subject matter data* respectively. Geography covers Countries, States, Counties, Cities, and so on. While Time addresses Years, Months, Weeks, Hours, and Minutes. Analysts then interact with the data in these tools, build dashboards, and visualizations for publication on federal websites, executive’s mobile devices and to be shared with other users or the general American public.

An increasing number of organizations have already done, or are currently planning to proceed with data streaming and migration tasks. However, most people underestimated the complexity of data streaming. According to the survey conducted by Howard and Potter, only 16% of migration projects were finished on time and on budget [36–38]. Well-

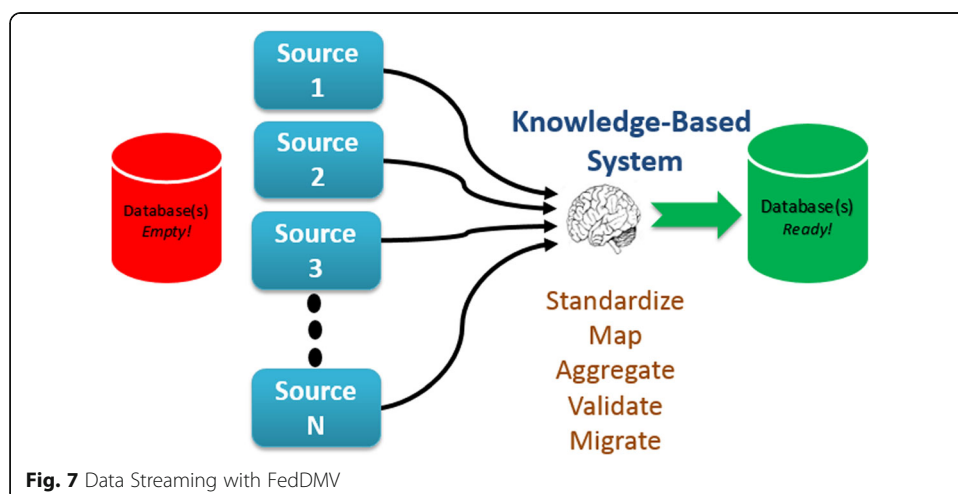


Fig. 7 Data Streaming with FedDMV

defined and practically proved models and methodologies have been sought by both the research and industry communities. Morris, in his book, the author discussed streaming data from legacy systems to new destinations, from planning, organizing, to implementing data migration projects [37, 39–41]. The book was written for readers including organization executives and data migration practitioners, and it mainly presented methodologies and pitfalls. All of them are very general and from the perspective of strategies without any models or specific technical description. In the paper of Wu et al. [42] though, the Butterfly Methodology was proposed to solve the legacy system streaming and migration problem. It includes six major phases, each with a set of independent activities. The paper provided a comprehensive model of migration of a whole system (and mainly targeting the data of a system), instead of focusing on data migration only. The authors also stated that they need more practical experiments to explore the relationships between system variables. Furthermore, the papers of Haller, Thalheim and Wang focused on the theoretical perspective of data migration [43, 44]. The authors investigated the connection between two types of models: abstract model and concrete model. Then, they re-defined the Extract-Transform-Load (ETL) process in data streaming, and discussed two types: property-preserving transformation and property-enhancing transformation. Also from the theoretical perspective, Spivak considered data migration with mathematical concepts [45]. In the book by Anavi-Chaput et al. [40], the authors discussed the process of migrating PeopleSoft Applications from Oracle to DB2 [46, 47] with highly detailed steps and scripts. These services include Transportable Tablespaces (TTS), Transportable Databases (TDB), Data Pump, and Recovery Manager (RMAN). Validation in migration is a very important aspect as well, studies looked at migration from a testing/validation perspective [48–50]. Similarly, several other online articles written by Scheier [51], Burry [52], Klazema [53], and Levine [54] discussed the average process and common errors of data migration, but none of the methods injected validation into streaming or paid attention to its risks and increased importance. Moreover, how can analysts trust this streamed data if it is not validated? FedDMV uses a method called ART (Association Driven Testing) or ADT (Association Driven Testing) - both ART and ADT are used interchangeably - to validate the streamed data system, and find potential errors. A sample validation SQL code is in Fig. 8 below, and streaming SQL sample in Fig. 9. However, in big data, finding an error is difficult, due to size and variety of data; so how does FedDMV determine the location of these errors? The next section introduces FedDMV's validation and addresses all the mentioned challenges.

Federal data validation

As it is well known, if the data migrated is not sufficiently validated, it would be extremely risky to the owners of the data in the future. Data validation is not used only to compare data between the origin and the destination, but also to ensure the *semantical* correctness, completeness, consistency, and the interoperability [48]. Especially, nowadays, simple databases have been growing tremendously into “big data”, many business intelligence and decision making highly rely on “data”, lower data quality may result in unsuccessful business strategies, which will further bring severe financial losses to any organization. Therefore, migrating data is always risky, and should be planned, implemented, and validated seriously. To mitigate the risks of the data anomalies, the research community proposed different state-of-the-art models and solutions. Researchers looked into

```

ALTER PROCEDURE [CoSD].[ValidationEngine-SummaryStatistics]
AS
BEGIN
-----***Datavalues Summary Statistics***-----
;with cteBase as (
    Select RowNr=Row_Number() over (Partition By DV.ERSDataValues_ERSC
    ,GrpByID = DV.ERSDataValues_ERSCCommodity_ID
    ,datavaluesyear = TD.ERSTimeDimension_Year
    ,CDS.ERSCommodity_SourceSeriesID_LongDesc + '+' + SL.ERS
    ,Measure = DV.ERSDataValues_AttributeValue
FROM [CoSD].[ERSDataValues] DV INNER JOIN [CoSD].[ERSCommodityDataSerie
INNER JOIN [CoSD].[ERSTimeDimension_LU] TD ON DV.ERSDataValues_ERSTimeD
INNER JOIN [CoSD].[ERSStatisticType_LU] SL ON CDS.ERSCommodity_ERStati
INNER JOIN [CoSD].[ERSProdPractice_LU] PL ON CDS.ERSCommodity_ERSProdPr
INNER JOIN [CoSD].[ERSUnit_LU] UL ON UL.ERSUnit_ID = DV.ERSDataValues_E
WHERE ERSDataValues_DataRowLifecyclePhaseID IN (1,4)
)
,cteMedian as (Select A.GrpByID,Measure From cteBase A Join (Select
,cteMode as (Select * from (Select RowNr=Row_Number() over (Parti
Select A.GrpByID AS 'CommodityID'
    ,A.LongDescription AS 'Commodity Description'
    ,RecordCount = Count(*)
    ,StartYear = MIN(A.datavaluesyear)
    ,EndYear = MAX(A.datavaluesyear)
    ,Minimum = Min(A.Measure)
    ,Maximum = Max(A.Measure)
    ,Mean = Avg(A.Measure)
    ,Median = Max(B.Measure)
    ,Mode = Max(C.Measure)
    ,StdDev = STDEV(A.Measure)
From cteBase A
LEFT Join cteMedian B on A.GrpByID=B.GrpByID
LEFT Join cteMode C on A.GrpByID=C.GrpByID
Group By A.GrpByID,A.LongDescription
Order By A.GrpByID

```

Fig. 8 Data Validation SQL Example

```

Create PROCEDURE [CoSD].[sp_Data_Streaming]
AS
BEGIN
    SET NOCOUNT ON;
    DECLARE @MySource varchar(50) = 'WASDE'
    DECLARE @C TABLE (Operations tinyint) --Temporay table variable for storing statistics updates
    ;

    WITH pou_liv AS
    (
    Select * from [AnimalProductsCoSD].[Wasde].[AP_GetLivestocks]
    UNION ALL
    Select * from [AnimalProductsCoSD].[Wasde].[AP_GetPoultry]
    )

    -- Join stream knowledge table from the master CoSD and dimensions from the CoSD
    , StreamThis AS
    (
    Select
    1.ERSSource_ID,
    1.ERSSource_Desc,
    1.ERSCommodity_ID,
    s.*
    FROM MasterCoSD.Wasde.KB s Left Outer join [dbo].[View_DataSeriesDimension] l on
    1.ERSSource_Desc = @MySource
    and s.ERSSector_Desc = 1.ERSSector_Desc
    and s.ERSGroup_Desc = 1.ERSGroup_Desc
    and s.ERSPhysicalAttribute_Desc = 1.ERSPhysicalAttribute_Desc
    and s.ERSCommoditySubCommodity_Desc = 1.ERSCommoditySubCommodity_Desc
    and s.ERSStatisticType_Attribute = 1.ERSStatisticType_Attribute
    and s.ERSProdPractice_Desc = 1.ERSProdPractice_Desc
    and s.ERSUtilPractice_Desc = 1.ERSUtilPractice_Desc
    where s.Enable = 1 --Just select tables that is enables
    )

```

Fig. 9 Data Streaming SQL Example from the US Department of Agriculture

common errors and risks of data migration that should be taken into consideration, but didn't provide any solutions [46, 49, 50, 52–55]. Different organizations may have different requirements about how to evaluate the data quality, and there is not a single universal technique that can solve it. Woodall et al. for example, proposed a hybrid approach of multiple methods to dynamically assess the data quality [56, 57]. However, based on the recent study by NIST [58, 59], the data error is not the challenge, it is actually the ability to define *the locations of errors*. That is listed as the most time-consuming activity of data validation. In the study, the NIST researchers compiled a vast number of software and data analytics projects and reached the following conclusion:

“If the location of bugs can be made more precise, both the calendar time and resource requirements of testing can be reduced. Modern data and software products typically contain millions of lines of code. Precisely locating the source of bugs in that code can be very resource consuming.”

Literature therefore, has a gap in error allocation; no or few AI-driven method was introduced to help analysts and engineers pin-point to errors that would affect the overall health of a data system. Some methods touched on applying AI logic to software engineering, and some others developed means to locate errors that are not based on a solid AI approach. FedDMV uses association rules for that purpose, ART is presented next.

Association rules validation and testing

Association Rules (AR) is one of the most commonplace data analytical models [60]. AR are intended to identify patterns of the type: “Action B often comes after action A , and is followed by Action C ”. One of the more well-studied problems in data mining is the search for AR in market basket data, or mostly referred to Market Basket Analysis (MBA). MBA allows businesses to realize which commodities are bought together, and therefore, they put them close to each other on store shelves. The outcome of the AR model is a set of rules. These rules include consequents ($B \rightarrow C$), antecedents ($A \rightarrow B$), support, and confidence of the rules. Confidence measures the reliability of the inference made by a rule. For a given rule $B \rightarrow C$, the higher the confidence, the more likely it is for C to be present in transactions that contain B . Confidence also provides an estimate of the conditional probability of C given B .

Support is an important measure as well, because a rule that has very low support may occur simply by chance. Association Rules outputs should be interpreted with thoughtfulness. The inference made by an association rule does not necessarily imply causality. Instead, it suggests a strong co-occurrence relationship between items in the antecedent and consequent of the rule.

Support and confidence are bound with these rules:

$$\text{Support} = (X \cup Y) \times \text{count of tests} \div n$$

$$\text{Confidence} = (X \cup Y) \times \text{count of tests} \div X \times \text{count of tests}$$

In this paper, ART aids in locating errors in data after streaming, by looking at the consequents of existing data errors [60]. For example, a data error occurred in module 1 function 2 (M1F2) of streaming; due to the module's importance, this module effects other modules in the system. Due to the AR statistical model, M7F10, M10F11 are the

most two functions that historically effect errors in M1F2 (M7F10, M10F11 → M1F2). While M6F1 is the function that will be effected due to the error that exist in M1F2. This M1F2 → M6F1 rule is what the engineer is looking for. However, they need to confirm that the rule has high confidence and acceptable support, they look at the confidence, and realize that its 0.805, which is considered very high (a confidence of 0.75 and up is considered *very high*, 0.50–0.75 is *high*, 0.30–0.50 is *low*, 0.30 and lower is *very low*). After the current testing iteration, the engineering team should perform heavy testing on the location that the ART engine recommended: M6F1. The testing needs to be done before the actual release of the data to the federal employee.

- 1 Data Collection: In order to increase the accuracy of any data mining model, a fair amount of high quality data required. More data helps the analysts and the engineers get a better quality model that could be more insightful.
- 2 Model Development and Data Training: In this stage, the AR model is built and the data is trained.
- 3 AR Model Outputs: Outputs of the model are antecedents, consequents, confidence and support.
- 4 Sort all Predictions: Consequent predictions are sorted by confidence, top predictions with highest confidence results are then considered for testing with data streaming.
- 5 Federal intelligent testing: using the outcomes from step 3 and the predictions from step 4, testing is performed and focused on the system modules present in the AR model's consequents.

Federal data security

Database security is about preventing malicious users from accessing or modifying data stored in databases. Similar to the generic security concept, three key aspects determine how secured a database is: Confidentiality, Integrity, and Availability. Confidentiality refers to only disclosing data to authorized users. To verify a user's identity and control access to data, database management systems (DBMS) use different methods of authentication [61–63]. Integrity refers to protecting the database from unauthorized writing. Data stored in databases should not be modified improperly, (i.e. corrupted). At the age of "big data", protecting database is one of the critical missions. Access control was one of the earliest database security measures proposed and widely used. In most cases, access control models can be categorized into three classes: discretionary access control (DAC), mandatory access control (MAC), and role-based access control (RBAC) [62–65] – this is the method adopted by FedDMV. The reason for that is based on the notion that federal data control is dependent on employees' roles, access levels, and aggregations. The next section introduces the experimental work that was performed with FedDMV.

FedDMV experimental studies

This section introduces the main three experimental works of this paper, the first is through comparing a big data analytics system built using FedDMV to another

built through traditional relational database; the second by surveying feedback on FedDMV’s tool from federal employees and analysts. The third is a test of ART, FedDMV’s validation method. The subsections present pros and cons of the two compared processes, outcomes of the comparison, results of the survey, ART and the lessons learnt.

Setup for the experiments

FedDMV was deployed to develop a data system at a US federal agency. The steps presented in the previous section are applied to a new data system development process. This section compares FedDMV to the *traditional* process that most federal agencies are currently using. A big data system was built using FedDMV, and another one (with similar size) was built through Traditional Relational Database approaches (that is referred to as TRDb here). The two process’s steps are presented in Table 2. The federal agency’s identity and information are not presented in this paper due to government data restrictions, however, the project that is managed by George Mason University is funded by the United States Department of Agriculture (USDA). For further details, please consult with the authors.

Data was collected to run the experiment a total of 633 rows was assembled. The system under test has 11 modules, with an average of 58 functions. Due to proprietary reasons, the nature and details of the system can’t be exposed.

The FedDMV process was performed to build a federal data system at USDA, at each stage, certain decisions were taken. The code shown in Fig. 9 creates a cursor that goes through all data and migrates it to tables in a destination database. Such procedures could be processed using SQL engines. The federal employees require friendly graphical user interfaces to be able to interact with the data. That includes the ability to upload, update, delete, and retrieve data from the FedDMV system. A tool is part of the process for data management. Developing a tool goes through all the steps of software development, coding, testing, and deployment. An example screenshot of the tool is shown in Fig. 10 (the tool provides a GUI for analysts to interact, stream and validate the data).

As for ART, a sample data that is used as part of the experiments is shown in Fig. 11 (after sorting rows by confidence). The figure shows the errors that previously occurred in Column A, the Antecedents in Column B, the Consequents in Column C (which what the engineer look for at each stage). The figure also shows Confidence, Lift and

Table 2 FedDMV Steps vs. TRDb steps

Process/Step	FedDMV	TRDb
Standards Creation	Standards are created based on the data streams and the data sources. All standards are saved in a knowledge base.	Lookups created based on informal conversations with federal employees
Data Management	Automated data management using a software tool.	Manual migrations (Excel and SSIS)
Data Validation	Using an automated Association Rules Driven approach to validate the data extensively.	Manual data validation and verifications practices
FedDMV Tool Development and Testing	A tool that gives federal employees access to the system through a GUI.	No such tool available
Data Security and Federal Restrictions Deployment	Well documented security and privacy routines.	Routines created as required

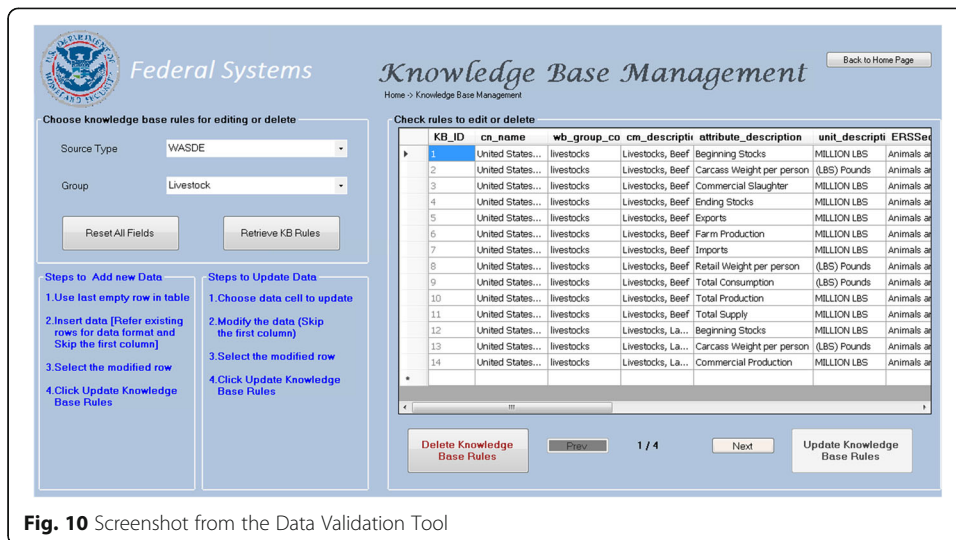


Fig. 10 Screenshot from the Data Validation Tool

the Support in columns D, E, and F. This data was collected and generated using ART throughout the development and testing process. The results of the three experiments are presented next.

Experimental results and conclusions

This section presents the experimental results of the three studies, pros and cons of FedDMV and TRDb, conclusions and future work.

Results

Time consumption and lessons learnt were collected from both processes, time consumption for FedDMV is significantly less than that of TRDb (45 vs. 81 months). Results for FedDMV are shown in Table 3 below, results for TRDb are in Table 4. Furthermore, advantages of FedDMV and disadvantages of TRDb are in Table 5.

For data management models, it is very important to include the end-users and collect their feedback. Therefore, we presented the FedDMV system to 350 governmental employees and analysts and collected their feedback on FedDMV's usability. The feedback was

Error in Function	Antecedent Functions	Consequent Function	Confidence	Lift	Support
M1F1	M7F21	M8F14	1	211.4091928	2.16E-03
M1F2	M7F1, M10F11	M6F1	0.804856115	6.314847165	4.75E-03
M1F3	M1F1	M1F35	0.676	46.86693088	2.69E-03
M1F4	M7F1, M6F1	M10F11	0.656639765	11.05106265	4.75E-03
M1F5	M10F11, M1F1	M6F1	0.599550056	4.70402958	2.83E-03
M1F6	M10F11, M5F3	M6F1	0.584745763	4.587876085	1.83E-03
M1F7	M1F1, M5F1	M6F1	0.56443299	4.428503428	1.16E-03
M1F8	M1F1, M5F1	M10F11	0.551546392	9.28237072	1.13E-03
M1F9	M10F11, M9F1	M6F1	0.551272167	4.32524491	3.79E-03
M1F10	M10F11	M6F1	0.544090056	4.268894135	1.54E-03
M1F11	M5F1, M6F1	M10F11	0.539968652	9.087520615	3.65E-03
M1F12	M6F1, M7F2	M2F3	0.535236396	2.457830659	3.18E-03
M1F13	M1F1, M9F1	M6F1	0.518382353	4.067193217	1.50E-03
M1F14	M10F11	M6F1	0.50965251	3.998699451	1.40E-03
M1F15	M10F11, M7F17	M6F1	0.501694915	3.936264699	2.35E-03

Fig. 11 Example Use Cases and Associations from FedDMV

Table 3 FedDMV Time Consumption

Process/Step	FedDMV
Standards Creation	2 months
Data Management	5 months
Data Validation (ART)	7 month
FedDMV Tool Development and Testing	22 months
Data Security and Federal Restrictions Deployment	9 months
Total Time	45 months

mostly positive. Feedback was categorized into the following five groups: FedDMV’s tool is user friendly, FedDMV is an inevitable change (willing to accept it), FedDMV is easy to follow, haven’t used the tool, dislike FedDMV and its tool. See results illustrated in Fig. 12. We noticed that new federal hires are usually early adopters of such data technologies, and are excited to see government systems move in the right direction and adopt new technologies. Nonetheless, many employees whom been with a federal agency for 15 years or more are very strict in terms of resisting change and refusing to change their ‘old’ ways.

For example, 9% of the analysts gave us negative feedback that us relevant to the notion that they don’t see the need to move away from more manual processes. However, 71% gave great or good feedback, and 20% thought that this change is necessary and eventually inevitable.

The results of the tool usability survey are considered to be very good in terms of use acceptance. FedDMV is being used at multiple federal agencies now, which is considered another proof of the success of the model and its tools.

As for ART, the results were collected after performing 11 iterations of streaming and testing. The results of the experiment are as follows: the riskiest module across all iterations is M6F1 (more than 150 occurrences); Refer to Fig. 13. This module is expected to have the highest number of errors. However, the module that the model has highest confidence that it will have errors is M8F14 (the only module with a confidence of 1). Refer to Fig. 14. As it is evident from the experimental results, many functions that are directly affected by errors in the data system are the first function in the module. As Fig. 13 suggests, M6F1, M3F1, M5F1 and M7F1 are modules with evident high risk. The AR model’s average confidence of the data system is 0.55011; therefore, most recommendations in this model are dependable as testing candidates. The data system module used for the experiment was released on

Table 4 TRDb Time Consumption

Process/Step	TRDb
Requirements Gathering and Variables’ Understanding	8 months
Database Design	7 months
Building Database Tables	10 month
Data Migration and Management	32 months
Business Rules Development	18 months
Giving Federal Employees Access to Data System	6 months
Total Time	81 months

Table 5 FedDMV vs. TRDb (Pros and Cons)

FedDMV	TRDb
Easy data management of centralized DB.	No "single version of the truth".
Better data validation practices that are driven by intelligent methods (association rules).	Manual validation.
No more manual manipulations with data and tables, the analysts use the FedDMV tool (leads to higher security).	updates and maintenance difficulties.
Resilient data integration and standardization routines.	Data not comprehensively structured in tables. Difficult publishing and sharing.
Less manual work by engineers and federal employees. Automated data streaming routines available.	Data overlap/redundancy between different databases – results in many inconsistencies. No automated data streaming routines.
Easier sharing and publishing. Role-based data security.	Uses hand entered data that increases the risk of errors. Data security constantly compromised.
Access to advanced analytical capabilities.	TRDb lacks tools that allow access to analytics.

time, and on budget and is currently in use by federal team. We aim to provide more experiments to ART in the future. Conclusions and future work are presented next.

Discussion and Conclusion

The 4-step model in this paper aims to guide the deployment process of a federal big data analytical system. For any data model, the following four factors are used to evaluate its efficiency and effectiveness:

- 1 Resource consumption rates: is the process expensive? How many hours? Manpower?
- 2 Complexity: how difficult is the model to follow?
- 3 Practicality: is the model merely theoretical or is it a good fit for real world projects?
- 4 Validity: ensuring that the data and the processes are valid

In the experimental study of this paper, FedDMV is compared with TRDb, a traditional process for building data systems. To assess the four points: in terms of resource consumption, FedDMV presented a significant improvement over traditional federal processes (45 months vs. 81 months), and was successful in delivering a big data

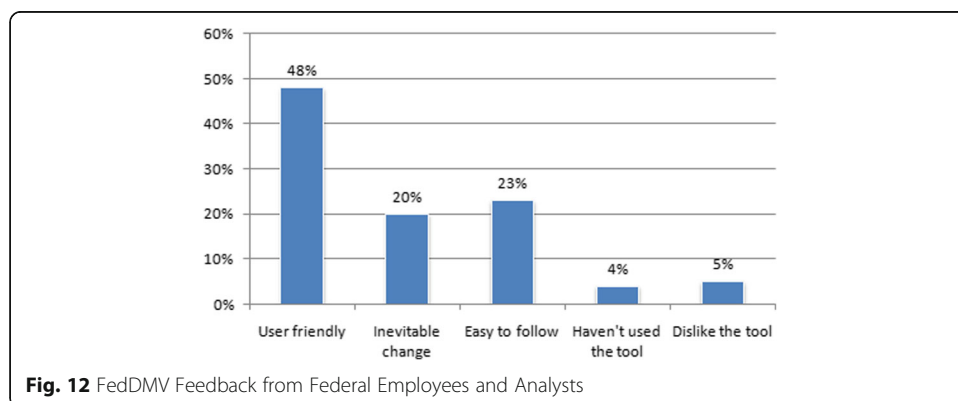
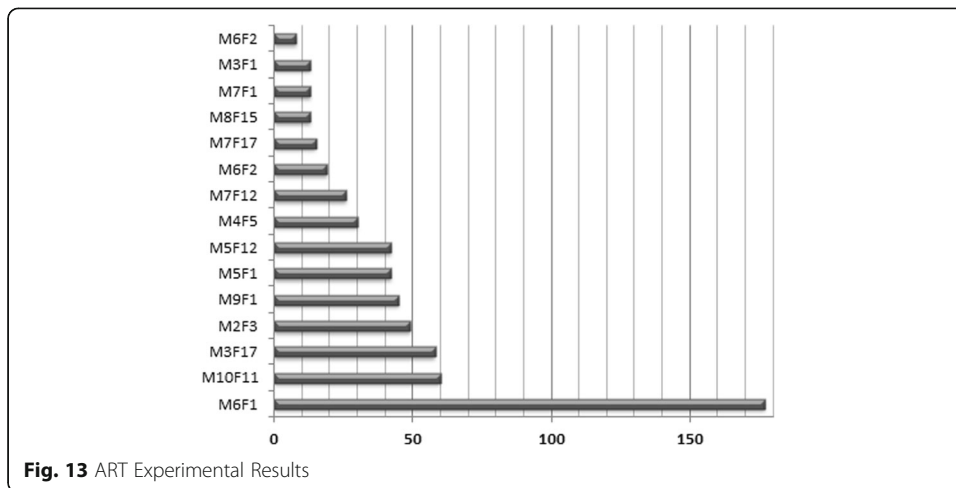


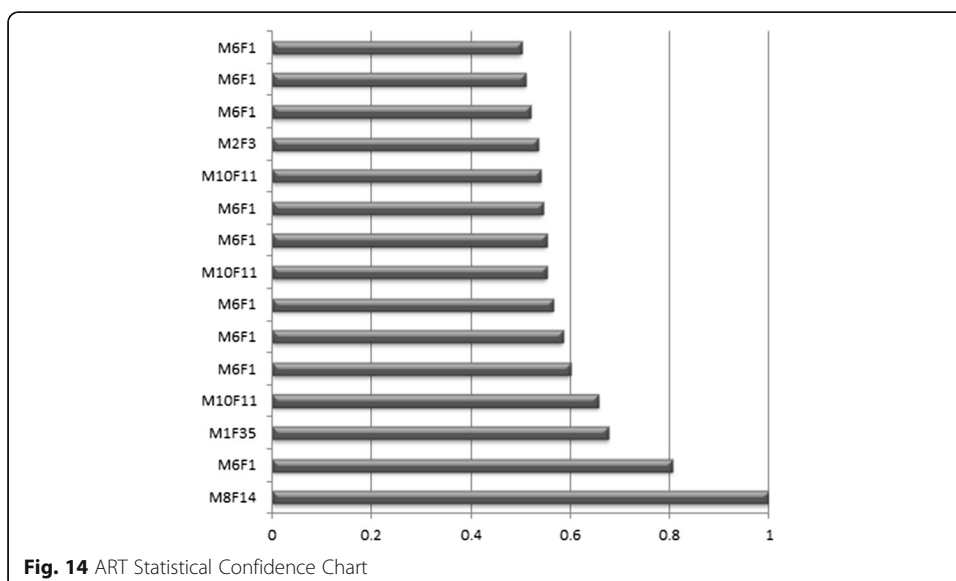
Fig. 12 FedDMV Feedback from Federal Employees and Analysts



system at a federal agency in a timely manner. In terms of complexity and practicality, FedDMV has 4 clear steps, and after working with multiple engineers at the federal agency, there is a general consent that FedDMV is easy to follow (the tool survey shows that as well). As for validity, FedDMV focuses of validation through ART.

Future work of our research includes the following:

1. Perform more experiments for ART and general quality assurance of data systems.
2. Deploy big data systems with other federal agencies. That will put FedDMV to test with different types of data, and different types of processes.
3. Deploy FedDMV with smaller types of systems and evaluate its feasibility with that.
4. Compare FedDMV to more data analytics models, in terms of time consumption, usability and applicability - such as PACER and BIRT.
5. Provide more software tools for managing the FedDMV process, and aid in project planning, tracking resources, and monitoring the 4 steps of the process.



FedDMV focuses on major federal big data aspects such as validation, streaming, security, and automation. In this age of big data, FedDMV is introduced as an effective and efficient 4-step model to follow, and is a very strong candidate for federal agencies that aim to develop new data systems.

Abbreviations

ACA: Affordable Care Act; AGA: Association of Government Accountants; DBMS: Database Management Systems; FedDMV: Federal Data Management and Validation; FMSS: Federal Mission Software Solution; GAO: Government Accountability Office; MTED: Market and Trade Economics Division; NASS: National Agricultural Statistics Service; PACER: Public Courts Online Access System; TRDb: Traditional Relational Database; USDA: United States Department of Agriculture

Acknowledgments

The authors would like to convey their thanks to the staff of the United States Department of Agriculture - Economic Research Services (ERS); especially the Applications Development Branch (ADB) and the Market and Trade Economics Division (MTED). The views and opinions expressed in this research paper are those of the authors and do not reflect the official policy or position of any agency or department of the U.S. government. Additionally, an acknowledgement goes to the following George Mason University students for their help and hard work with the federal projects: Gowtham Ramamoorthy, Manish Dashora and Samantha Dcosta.

Funding

Not applicable.

Availability of data and materials

The actual Data is not shared because this is federal data and parts of the data set are private or confidential, therefore, we are not allowed to share it. However, to get a sample data set visit: <http://ers.usda.gov/data-products.aspx>. All the software tools used in the experiments and as part of FedDMV are available via the online sharing website (Dropbox), upon request from the author.

Authors' contributions

FB worked on the overall design and development of the FedDMV model and the associated system, executed the experiments, collected data for the comparison study, and interacted with the federal employees on frequent basis. FB wrote multiple sections of this manuscript and provided the illustrations. RY worked with the federal agency management to setup the experiments. RY wrote multiple sections of the paper and provided multiple insights. LD developed the tool, and executed some of the experiments. LD also wrote some parts of the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

¹College of Science, George Mason University (GMU), Fairfax, VA, USA. ²Department of Computer Science, George Mason University (GMU), Fairfax, VA, USA.

Received: 3 May 2016 Accepted: 21 October 2016

Published online: 10 January 2017

References

1. Turing AM. Computing machinery and intelligence. *Mind*. 1950;49:433–60.
2. McCarthy J. Programs with common sense. In: *Proceedings of the Symposium on Mechanisation of Thought Processes*, vol. 1. London: Her Majesty's Stationery Office; 1958. p. 77–84.
3. Koh HC, Gerald T. Data Mining Applications in Healthcare. *J Healthc Inf Manage*. 19. 2.
4. Jing L. Data Mining Applications in Higher Education, SPSS Executive Report. 2004.
5. Book: Industry Applications of Data Mining. Chapter 8, 1999.
6. Harmann M. The Role of Artificial Intelligence in Software Engineering. Published report at the Crest Center, University College London.
7. An Immediate Release by the Office of Science and Technology Policy. Executive Office of the President. The Big Data Initiative. 2012.
8. The White House Big Government Initiative: <http://www.whitehouse.gov/Open/>.
9. White Paper by Accenture. Accenture Federal Services, Federal Analytics and Big Data. 2012.
10. Letter from James Madison to W.T. Barry (Aug. 4, 1822), reprinted in *The Writings of James Madison* (Gaillard Hunt). 11.
11. A White Paper by CC Pace Systems. Agile in the Federal Government. 2014.
12. Batarseh FA. Incremental Lifecycle Validation of Knowledge-Based Systems through CommonKADS. PhD Dissertation Registered at the University of Central Florida and the Library of Congress; 2012.
13. DePillis L. The Way Government does Tech is Outdated and Risky. A Report Published at the Washington Post; 2013.

14. Smith M, Cohen T, A CNN Report on the ACA Website. Problems with Health Website. 2013.
15. United States Government Accountability Office (GAO). Software Development - Effective Practices and Federal Challenges in Applying Agile Methods, GAO-12-681, gao.gov. 2012.
16. Batarseh F, Gonzalez A. Predicting Failures in Contextual Software Development through Data Analytics. Proc Springer Softw Qual J. 2015.
17. Ming-Tun Yu H. Designing Software To Shape Open Government Policy. A Dissertation Presented to the Faculty of Princeton University in Candidacy for the Degree of Doctor of Philosophy. The Department of Computer Science, Princeton University; 2012.
18. Hahn R. Government Policy toward Open Source Software: An Overview. Chapter 1. Published at the Brookings Institute.
19. National Agricultural Statistics Service: <http://quickstats.nass.usda.gov/>.
20. City of Chicago Data Portal: <http://data.cityofchicago.org>.
21. San Francisco Governmental Data Website: <http://data.sfgov.org>.
22. Swish-Data - Data Performance Architects, Hadoop Uses Cases: Big Data for the Government, <http://www.swishdata.com/index.php/blog/article/hadoop-use-cases-big-data-for-the-government>.
23. MapR: Big Data and Apache Hadoop for Government: <http://www.mapr.com/solutions/industry/big-data-and-apache-hadoop-government>.
24. Horton Works by Apache: <http://hortonworks.com/industry/government/>.
25. Microsoft's SQL Server: <http://www.microsoft.com/SQLserver>.
26. SAS: http://www.sas.com/en_us/home.html.
27. Tableau: <http://www.tableau.com/>.
28. Batarseh F, Gonzalez AJ. Incremental Lifecycle Validation of Knowledge-Based Systems through CommonKADS. Published at the IEEE Transactions on Systems, Man and Cybernetics (TSMC-A); 2012.
29. Batarseh F. Context-Driven Testing. In: Context in Computing: A Cross-Disciplinary Approach for Modeling the Real World. Patrick Brezillon, Editor (Sorbonne University- Paris VI). Springer Verlag; 2015. ISBN: 978-1-4939-1886-7.
30. Scacchi W. Process Models in Software Engineering. Institute for Software Research, Encyclopedia of Software Engineering. 2nd Ed. John Wiley and Sons; 2001.
31. Gartner's Magic Quadrant: <http://www.gartner.com>.
32. A Report by Actuate. Reporting and Data Analytics for Federal Applications. Actuate BIRT; 2014
33. Report S. Federal Mission Software Solution. 2014.
34. Information-builders. Predictive Analytics for Federal Government. New York: Product Brochure; 2011.
35. Corporate Partner Advisory Group, Research Series. Leveraging Data Analytics in Federal Organizations, Report no. 30. 2012.
36. Howard P, Potter C. Data Migration in the Global 2000 - Research, Forecasts and Survey Results. 2007. p. 29.
37. Morris J. Practical Data Migration. 3rd ed. Swindon: British Informatics Society Ltd; 2006.
38. Inmon W, Linstedt D. Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault. Morgan Kaufmann; 2014.
39. A Report by IBM. The Hidden Costs of Data Migration - Strategies for Reducing Costs and Quickly Achieving Return on Investment. 2007.
40. Wei B, Chen T. Verifying Data Migration Correctness: The Checksum Principle. RTI Press; 2014. OP-0019-1403.
41. Department of Education, Office of Federal Student Aid. Data Migration Roadmap: A Best Practice Summary, Version 1.0. 2007.
42. Wu B, Lawless D, Bisbal J, Richardson R, Grimson J, Wade V, Sullivan D. The Butterfly Methodology: A Gateway-free Approach for Migrating Legacy Information Systems, Proceedings of the 3rd IEEE International Conference on Engineering of Complex Computer Systems (ICECCS97). Como: Institute of Electrical and Electronics Engineers; 1997. p. 200-5.
43. Haller K, Matthes F, Schulz C. Testing & Quality Assurance in Data Migration Projects, Proceedings of the 27th IEEE International Conference on Software Maintenance ICSM. 2011.
44. Thalheim B, Wang Q. Data Migration: A Theoretical Perspective. Data Knowl Eng. 2013;87:260-78.
45. Spivak D. Functional data migration. Inf Computing. 2012;217:31-51.
46. Oracle's Document: Move to Oracle Database with Oracle SQL Developer Migrations.
47. Oracle White Paper. Migrating Oracle Databases. 2014.
48. Haller K, Matthes F, Schulz C. Testing & Quality Assurance in Data Migration Projects. In: 27th IEEE International Conference on Software Maintenance ICSM. 2011.
49. Fine L, Keogh B, Cretin S, Orlando M, Gould M. How to Evaluate and Improve the Quality and Credibility of an Outcomes Database: Validation and Feedback Study on the UK Cardiac Surgery Experience. BMJ. 2003;326(7379):25-8.
50. Flamos A, Doukas H, Psarras J. Data Validation Platform for the Sophisticated Monitoring and Communication of the Energy Technology Sector. Renew Energy. 2010;35(5):931-5.
51. Scheier R. Data Migration Strategies and Best Practices. An Article Published at TechTarget.
52. Burry C, Mancusi D. How to Plan for Data Migration, Proceedings of the Advanced Information Systems Engineering: 21st International Conference. 2004.
53. Klazema A. Data Migration Strategies for Businesses. <http://blog.udemy.com/data-migration-strategies/>. An Udemy Blog.
54. Levine R, A web-report by Alta Flux Corporation. Data Migration Strategies. 2013.
55. Anavi-Chaput V, Arrell K, Baisden J, Corrhons R, Fallon D, Siegmund L, Sokolof N. Planning for a Migration of PeopleSoft 7.5 from Oracle/UNIX to DB2 for OS/390. Poughkeepsie; 2000. p. 148.
56. Woodall P, Borek A, Parlikad A. Data quality assessment: The Hybrid Approach. Info Manag. 2013;50(7):369-82.
57. A Report by IBM. Best practices for data migration - Methodologies for assessing, planning, moving and validating data migration. Somers; 2009. p. 16.
58. National Institute of Standards and Technology: WWW.NIST.GOV.
59. Planning Report for NIST (US Department of Commerce). The Economic Impacts of Inadequate Infrastructure for Software Testing. 2002.
60. Association Analysis: Basic Concepts and Algorithms, Introduction to Data Mining, Chapter 6, pp. 327-414, by: Pang-Ning tan, Michael Steinbach, and Vipinkumar.

61. Oracle's Database Concepts Guide: From the Oracle Help Center.
62. De Capitani di Vimercati S, Foresti S, Samarati P. Recent Advances in Access Control. In: Handbook of Database Security. 2008; pp. 1–26.
63. Bertino E, Jajodia S, Samarati P. Database security: Research and practice. Information Systems. 1995.
64. Hore B, Mehrotra S, Hacigümüş H. Managing and Querying Encrypted Data. In: Handbook of Database Security. 2008. p. 1–26.
65. Oracle's Database Advanced Security Administrator's Guide: From the Oracle Help Center.

Submit your next manuscript to BioMed Central
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

