A qualitative readiness-requirements assessment model for enterprise big-data infrastructure investment

Conference Paper · May 2014

DOI: 10.1117/12.2050605

CITATIONS
12

READS
2,310

4 authors, including:

Mohammed Olama
Oak Ridge National Laboratory
155 PUBLICATIONS **1,242** CITATIONS

SEE PROFILE

Sreenivas Rangan Sukumar
Oak Ridge National Laboratory
74 PUBLICATIONS **758** CITATIONS

SEE PROFILE

James J. Nutaro
Oak Ridge National Laboratory
92 PUBLICATIONS **1,204** CITATIONS

SEE PROFILE

A Qualitative Readiness-Requirements Assessment Model for Enterprise Big-Data Infrastructure Investment

Mohammed M. Olama*, Allen W. McNair, Sreenivas R. Sukumar, James J. Nutaro

Computational Sciences and Engineering Division, Oak Ridge National Laboratory
1 Bethel Valley Road, Oak Ridge, TN, USA 37831-6085

ABSTRACT

In the last three decades, there has been an exponential growth in the area of information technology providing the information processing needs of data-driven businesses in government, science, and private industry in the form of capturing, staging, integrating, conveying, analyzing, and transferring data that will help knowledge workers and decision makers make sound business decisions. Data integration across enterprise warehouses is one of the most challenging steps in the big data analytics strategy. Several levels of data integration have been identified across enterprise warehouses: data accessibility, common data platform, and consolidated data model. Each level of integration has its own set of complexities that requires a certain amount of time, budget, and resources to implement. Such levels of integration are designed to address the technical challenges inherent in consolidating the disparate data sources. In this paper, we present a methodology based on industry best practices to measure the readiness of an organization and its data sets against the different levels of data integration. We introduce a new Integration Level Model (ILM) tool, which is used for quantifying an organization and data system's readiness to share data at a certain level of data integration. It is based largely on the established and accepted framework provided in the Data Management Association (DAMA-DMBOK). It comprises several key data management functions and supporting activities, together with several environmental elements that describe and apply to each function. The proposed model scores the maturity of a system's data governance processes and provides a pragmatic methodology for evaluating integration risks. The higher the computed scores, the better managed the source data system and the greater the likelihood that the data system can be brought in at a higher level of integration.

Keywords: Big data, multi-agency data integration, data management, data warehouse

1. INTRODUCTION

Nowadays, we live in the "Big Data" era. Data is being generated, collected and analyzed at an unprecedented scale. This data is being collected in a large variety of domains, such as web text and documents, web logs, large-scale e-commerce, social networks, sensor networks, astronomy, genomics, medical records, surveillance, and others. A new study by IDC Digital Universe [1] predicts there will be 40 Zettabytes (ZB) of data on the planet by 2020, an amount that exceeds previous forecasts by 14%. To put that into a real world context, 40 ZB is equal to 57 times the number of all the grains of sand on all the beaches on earth.

The "Big Data" analytics is the ability to extract valuable information and discover unique insight from all this data in making data-driven decisions to alter all aspects of society. It enables organizations to improve decision making, resulting in the ability to take advantage of opportunities, minimize risks, and control costs. It also provides opportunities for businesses to drive innovation, improve productivity, detect fraud, enhance customer satisfaction, and increase profit margins.

Although many organizations recognize the value of big data, they have struggled to handle this new influx of data. The problem isn't information overload; it's the failure to harness, prioritize, and understand the data flowing in. This is why data integration is a critical step in the big data analytics strategy. Since the value of data explodes when it can be linked and fused with other data, addressing the big data integration challenge is critical to realizing the promise of Big Data.

*olamahussemm@ornl.gov; phone 1 865 574-8112; fax 1 865 576-0003

Next-Generation Analyst II, edited by Barbara D. Broome, David L. Hall, James Llinas,
Proc. of SPIE Vol. 9122, 91220E · © 2014 SPIE · CCC code: 0277-786X/14/$18
doi: 10.1117/12.2050605

Big data integration differs from traditional data integration in many dimensions such as volume (size), velocity (rate of data flow), variety (types of sources), viscosity (inertia to move), complexity and veracity (data quality). These challenges result from (i) the number of data sources, even for a single domain, has grown to be in the tens of thousands, (ii) many of the data sources are very dynamic, as large volumes of newly collected data are continuously made available, (iii) the data sources are extremely heterogeneous in their structure, with considerable variety even for conceptually similar entities, and (iv) the data sources are of widely differing quality, with significant differences in the coverage, accuracy and timeliness of data provided.

Multi-agency fully-integrated data is not recommended at initial stages because of its high cost and risk. Simpler levels of integration enable organizations to acquire experience with other organizations and uses of data residing in other agencies. Several levels of data integration have been identified across enterprise warehouses such as data access, data storage, harmonized data, data matching, and data grouping. Each level of integration has its own set of complexities that requires a certain amount of time, budget, and resources to implement. Such levels of integration are designed to address the technical challenges inherent in consolidating the disparate data sources. The choice of which data integration level is best depends upon several factors related to the organization readiness and its data sets. These factors are discussed in more detail in Section 3.

In this paper, we present a methodology based on industry best practices to measure the readiness of an organization and its data sets against the different levels of data integration. We also introduce a new Integration Level Model (ILM) tool, which is used for quantifying an organization and data system's readiness to share data at a certain level of data integration. It is based largely on the established and accepted framework provided in the Data Management Association (DAMA) International Data Management Body of Knowledge (DMBOK) [2]. It comprises several key data management functions and supporting activities, together with several environmental elements that describe and apply to each function. The proposed model scores the maturity of a system's data governance processes and provides a pragmatic methodology for evaluating integration risks. The higher the computed scores, the better managed the source data system and the greater the likelihood that the data system can be brought in at a higher level of integration.

The paper is structured as follows: Section 2 presents the most common data integration levels identified across enterprise warehouses. In Section 3, we introduce the key data management functions considered in the ILM tool that help in the assessment of the quality of an organization's data sets from technical and business perspectives. The developed ILM tool, which is used for quantifying an organization and data system's readiness to share data, is described in Section 4. Finally, Section 5 provides concluding remarks.

2. THE DIFFERENT LEVELS OF DATA INTEGRATION

It is recommended that multi-agency data integration begin at a simple level, thereby enabling organizations to acquire experience with other organizations and uses of data residing in other agencies. This experience will enable the accurate assessment of opportunities, problems, and potential solutions prior to embarking on a costly and potentially high risk program to create a fully-integrated data repository.

The three levels of data integration across enterprise warehouses considered in this paper are summarized as follows:

- **Level 1- Data Accessibility:** The first level provides a common interface to access the data. This is accomplished with a common portal or web service that provides access to as-is data from each agency/enterprise. Agencies can query against the other agencies' data residing in their "home" platforms. Central to this model is that data stays in place and that each agency has control of and is responsible for operating its data platform.
- **Level 2- Common Data Platform:** The second level creates a common platform for the integrated data. This is accomplished by locating data provided by each agency in a single, logical data platform. There are three models for the logical data platform: i) a federated model in which each of the data resides physically at each agency (as in the Level 1 above) but supports some type of uniform data access mechanism; ii) a physical model in which data is kept within a single data warehouse; and ii) a hybrid model of these models in which some agencies retain their data while permitting remote access to it and other agencies allow for replication of their data in the common data platform.

- **Level 3- Consolidated Data Model:** The third level of integration involves creating a consolidated data model, which will document and organize the data. The data model will attempt to organize the placement of the data into logical entities. It will accomplish this by grouping together data elements that are common across agencies and creating a common understanding of the data elements. This level will produce a harmonization of data across agencies. The harmonization of the data allows for a common data dictionary of data elements to be developed and agreed upon. At this level, all data is stored in a common data platform adheres to a common data model.

At each level of integration, specific technical, policy, and organizational issues can be identified. After surveying best practices in data management and enterprise information management, we were able to identify several key data management functions together with environmental elements applicable to each data management function. The identified data management functions and their associated environmental elements are presented in the next section.

3. IDENTIFYING KEY DATA MANAGEMENT FUNCTIONS

The main objective is to develop a tool for multi-agency data integration process. To accomplish this objective, we surveyed best practices in data management and enterprise information management. These best practices described various perspectives on enterprise data systems such as data management approaches [2], data governance maturity models [3], enterprise information management [4], business process maturity models [5], and the Institute of Electrical and Electronics Engineers (IEEE) systems definition and concept of operations [6].

After surveying best practices in data management and enterprise information management, we found out that they can be divided into two segments: technical and business. A technical segment defines the key supporting technologies for inclusion in a system. Initiation phases of projects have many unknowns; therefore, the technical segment should capture specific data, systems, analyses, and outputs that are critical to the enterprise. The technical segment can define specific measures that establish the entry and exit criteria for each level of integration. Examples of data elements included in the technical segment are: data architecture, data development, data quality, database operations, data security, data warehousing & business intelligence, and document management.

The business segment focuses on the desired outcomes related to data analytics, information sharing, required artifacts, and oversight activities. It defines the major business outcomes related to process change, deliverables, and special activities. Examples of data elements included in the business segment are: data governance, data management & control, data quality management, risk management, content management, and dispute & compliance resolution.

Both technical and business data management segments are considered in the development of the ILM tool. The 10 Data Management Functions considered in the model and their definitions are:

- Data Governance – planning, supervision and control over data management and use: The exercise of authority, control and shared decision-making (planning, monitoring and enforcement) over the management of data assets. Data Governance is high-level planning and control over data management.
- Data Architecture Management – as an integral part of the enterprise architecture: The development and maintenance of enterprise data architecture, within the context of all enterprise architecture, and its connection with the application system solutions and projects that implement enterprise architecture.
- Data Development – analysis, design, building, testing, deployment and maintenance: The data-focused activities within the system development lifecycle, including data modeling and data requirements analysis, design, implementation and maintenance of databases data-related solution components.
- Database Operations Management – support for structured physical data assets: Planning, control and support for structured data assets across the data lifecycle, from creation and acquisition through archival and purge.
- Data Security Management – ensuring privacy, confidentiality and appropriate access: Planning, implementation and control activities to ensure privacy and confidentiality and to prevent unauthorized and inappropriate data access, creation or change.

- Reference & Master Data Management – managing golden versions and replicas: Planning, implementation and control activities to ensure consistency of contextual data values with a "golden version" of these data values.
- Data Warehousing & Business Intelligence Management – enabling access to decision support data for reporting and analysis: Planning, implementation and control processes to provide decision support data and support knowledge workers engaged in reporting, query and analysis.
- Document & Content Management – storing, protecting, indexing and enabling access to data found in unstructured sources (electronic files and physical records): Planning, implementation and control activities to store, protect and access data found within electronic files and physical records (including text, graphics, image, audio, video)
- Meta Data Management – integrating, controlling and delivering Meta data: Planning, implementation and control activities to enable easy access to high quality, integrated Meta data.
- Data Quality Management – defining, monitoring and improving data quality: Planning, implementation and control activities that apply quality management techniques to measure, assess, improve and ensure the fitness of data for use.

Also there are Environmental Elements applicable to the data management functions that provide a logical and consistent way to describe each data management function. The environmental structure identifies three elements: process, technology and people. The category of process includes processes, deliverables, principles, and methods & techniques. The category of people includes roles & responsibilities, and organizational & cultural issues.

The 7 Environmental Elements considered in the ILM tool and their definitions are:

- Goals & Principles – The directional business goals of each function and the fundamental principles that guide performance of each function.
- Activities – Each function is further decomposed into lower level activities. Some activities are grouped into sub-functions. Activities can be further decomposed into tasks and steps.
- Deliverables – The information and physical databases and documents created as interim and final outputs of each function. Some are considered essential, some are generally recommended, and others are optional depending on circumstances.
- Roles and Responsibilities – The business and IT roles involved in performing and supervising the function and the specific responsibilities of each role in that function. Many roles will participate in multiple functions.
- Practices & Procedures – Common and popular methods and techniques used to perform the processes and produce the deliverables. May also include common conventions, best practice recommendations and alternative approaches without elaboration.
- Technology – Categories of supporting technology (primarily software tools), standards and protocols, product selection criteria and common learning curves.
- Organization and Culture – These issues might include: management metrics (such as measures of size, effort, time, cost, quality, effectiveness, productivity, success and business value), critical success factors, reporting structures, contracting strategies, budgeting and related resource allocation issues, teamwork and group dynamics, authority & empowerment, shared values & beliefs, expectations & attitudes, personal style & preference differences, cultural rites, rituals and symbols, organizational heritage, and change management recommendations

To quantify an organization's readiness to share data at any particular level of integration, we have developed the ILM tool. It will help an organization to select an integration strategy that is affordable, likely to succeed, and meets the technical requirements of its analysts and analysts within other organizations. This tool is based on the introduced data management functions and their associated environmental elements and it is discussed next.

4. THE INTEGRATION LEVEL MODELING TOOL

The focus of the Integration Level Modeling (ILM) tool is to quantify an organization and data system's readiness to share data at a given data integration level. We developed the ILM (implemented as an Excel Workbook) to provide a tool to score the maturity of a system's data governance processes. The higher the computed scores, the better managed the source data system and the greater the likelihood that the data system can be brought in at a higher level of integration.

The ILM tool is based on the DAMA-DMBOK's 10 key data management functions described in the previous section. For each of these data management functions, the model scores the maturity according to seven environmental elements also described in the previous section. To simplify the computation of the score for users of the ILM, the required data input is limited to drop-down selections in the Excel Workbook. The score computation consists of three steps: threshold setting, answering a questionnaire, and integration level calculation. These steps are described next.

4.1 Step 1: Threshold Setting

An enterprise management must configure a set of preferences as required maturity thresholds for the data management functions. The preference options are Low, Medium, or High – corresponding to the maturity required from a particular data management function to allow integration at a particular data integration level. For example, a Low setting in the Level 1 column suggests that the particular data management function does not need to be mature to allow data integration at Level 1. Though these preferences are adjustable according to the business requirements and engineering perspectives of the stakeholders, they should be standardized by the enterprise across the systems being evaluated. An example of the preference setting is illustrated in Fig. 1 below.

	Levels of Data Integration		
Data management Functions	Level 1: Data Accessibility	Level 2: Common Data Platform	Level 3: Consolidated Data Model
Data Governance	Medium	Medium	High
Data Architecture Management	Low	Medium	High
Data Development	Medium	Low	High
Database Operations Management	Medium	Medium	Low
Data Security Management	Medium	High	High
Reference & Master Data Management	Low	Medium	High
Data Warehousing & Business Intelligence Management	Low	Low	High
Document & Content Management	Low	Medium	Low
Meta Data Management	Low	Medium	High
Data Quality Management	High	Medium	Low

Fig. 1. Example of an enterprise management preference settings.

4.2 Step 2: Answering a Questionnaire for Each Data Management Function

Next, a team of business and technical staff from the organization must answer Yes/No questions for environmental elements of the operating system organization's 10 data management functions. This questionnaire resides within the ILM tool as a spreadsheet. The questions are used to determine and score the maturity of the data management function. Figure 2, below, gives an example of one such questionnaire. The maturity level for a particular data management function for an environmental element is the percentage of Yes answers for that element in the questionnaire.

1. Data Governance Questionnaire	Answers	Score (%)
Goals & Principles		50%
Are directional business goals of data governance well established within your organization?	Yes	
Are fundamental principles that guide performance of data governance well established within your organization?	No	
Activities		44%
Data Management Planning		44%
Does your organization regularly engage in identifying strategic enterprise data needs?	No	
Does your organization regularly engage in developing & maintaining the data strategy?	No	
Does your organization regularly engage in establishing the data management professional organizations?	Yes	
Does your organization regularly engage in identifying & appointing data stewards?	Yes	
Does your organization regularly engage in establishing data governance & stewardship organizations?	Yes	
Does your organization regularly engage in developing, reviewing & approving data policies, standards and procedures?	No	
Does your organization regularly engage in reviewing & approving data architecture?	No	
Does your organization regularly engage in planning and sponsoring data management projects & services?	No	
Does your organization regularly engage in estimating data asset value & associated data management costs?	Yes	
Data Management Supervision & Control		43%
Does your organization regularly engage in supervising the data management professional staff & organizations?	Yes	
Does your organization regularly engage in coordinating data governance activities?	No	
Does your organization regularly engage in managing & resolving data related issues?	No	
Does your organization regularly engage in monitoring & ensuring regulatory compliance?	No	
Does your organization regularly engage in monitoring conformance with data policies, standards and architecture?	No	
Does your organization regularly engage in overseeing data management projects & services?	Yes	
Does your organization regularly engage in communicating & promoting the value of data assets?	Yes	
Deliverables		50%
Does your organization regularly produce information and physical databases of data governance?	No	
Does your organization regularly produce reports and documents of data governance?	Yes	
Roles & Responsibilities		0%
Are there clearly defined business and IT roles involved in performing and supervising the data governance ?	No	
Are there specifically defined responsibilities of the main roles in the data governance?	No	

Fig. 2. Example of ILM Questionnaire.

4.3 Step 3: Recommended Integration Level Calculation

From the user-provided Yes/No responses to the questionnaire, the tool computes a "Yes", "No", or "Somewhat" assessment for the various environmental elements within a given data management function. These outcomes may be interpreted as follows:

- Yes – The organization has a mature capability with regards to the given data management function's environmental element.
- No – The organization has limited capability with regards to the given data management function's environmental element.
- Somewhat – The organization has a moderate capability with regards to the given data management function's environmental element.

Using this computed assessment (Yes, No, Somewhat), the model then determines the maturity of a given data management function. An organization is capable of supporting a given level of integration if its score in each functional area exceeds the minimum required threshold for that level.

The final output of this tool is an Integration Status Report (ISR), as illustrated in Fig. 3, which includes a recommended integration level. The higher the computed scores, the better managed the source data system and the greater the likelihood that the data system can be brought in at a higher level of integration. The colored cells in the example report show how mature the organization is with regards to specific data management activities as defined in the DAMA-DMBOK. The organization's scores in each functional area, the minimal required score, and the highest supported level of data integration are shown on the right side of the table.

An enterprise can use the ISR as a maturity model to show executives its organization's data-sharing readiness, and where targeted investments may improve the organization's posture to allow integration at a higher level.

	Environmental Elements							Organizational Capability	Required Capability for Levels of Integration			
Data Management Functions	Goals & Principles	Activities	Deliverables	Roles & Responsibilities	Technology	Practices & Techniques	Organization & Culture	Score	Lvl 1	Lvl 2	Lvl 3	Highest Supported Level of Integration
Data Governance	Somewhat	Somewhat	Somewhat	No	Somewhat	No	Somewhat	36%	30%	30%	60%	2
Data Architecture Management	Somewhat	Yes	Somewhat	No	Somewhat	Somewhat	Somewhat	50%	0%	30%	60%	
Data Development	No	No	No	No	Somewhat	Somewhat	Somewhat	21%	30%	0%	60%	
Database Operations Management	Somewhat	Somewhat	Somewhat	No	No	Somewhat	Somewhat	36%	30%	30%	0%	
Data Security Management	Yes	No	Somewhat	Somewhat	Yes	Yes	Somewhat	64%	30%	60%	60%	
Reference & Master Data Management	Yes	Somewhat	No	Somewhat	Yes	Somewhat	No	50%	0%	30%	60%	
Data Warehousing & Business Intelligence Management	Somewhat	No	No	No	No	No	Somewhat	14%	0%	0%	60%	
Document & Content Management	Yes	No	No	No	Somewhat	Somewhat	Somewhat	36%	0%	30%	0%	
Meta Data Management	Somewhat	Somewhat	Somewhat	No	Somewhat	Somewhat	Somewhat	43%	0%	30%	60%	
Data Quality Management	Somewhat	Somewhat	No	Yes	Somewhat	No	No	36%	60%	30%	0%	
							Average	39%	18%	27%	42%	

Fig. 3. Integration Status Report.

5. CONCLUSION

Based on our experience working with data assets both in the private industry and with the government, we conclude that the success in multi-agency data integration projects is a function of organization, personnel, technology and the characteristics of the data that is being brought together for analysis. These attributes are captured while developing the ILM tool, which provides a pragmatic methodology for evaluating integration risks. It quantifies an organization and data system's readiness to share data at a certain level of data integration, in addition to pointing out the fields/areas where targeted investments may improve the organization's posture to allow integration at a higher level. We believe that by employing the ILM tool, organizations will more quickly and efficiently be able to evaluate the risks associated with potential integration projects.

ACKNOWLEDGEMENTS

This paper has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

REFERENCES

[1] "Data to grow more quickly says IDC's Digital Universe study," Computer Weekly News, (12 December 2012). Available at: http://www.computerweekly.com/news/2240174381/Data-to-grow-more-quickly-says-IDCs-Digital-Universe-study.

[2] Mosley, M., "DAMA-DMBOK Functional Framework", Version 3.02, Data Management Association, (2009). Available at: http://dama.org/i4a/pages/index.cfm?pageid=3548.

[3] "Data Governance Part II: Maturity Models – A Path to Progress," National Association of State Chief Information Officers (NASCIO). (March 2009). Available at: http://www.nascio.org/publications/documents/NASCIO-DataGovernancePTII.pdf.

[4] Newman, D., and Logan, D., "Gartner Introduces the Enterprise Information Management (EIM) Maturity Model," Gartner Research, ID Number: G00160425, (2008).

[5] "Business Process Maturity Model," Object Management Group, Version 1.0, (2008). Available at: http://www.omg.org/spec/BPMM/1.0/PDF.

[6] "IEEE Guide for Information Technology—System Definition—Concept of Operations (ConOps) Document," Institute of Electrical and Electronics Engineers, Inc. (IEEE Std 1362-1998). (19 March 1998).