OPEN DATA CENTER ALLIANCE℠ USAGE MODEL: DATA MANAGEMENT FOR INFORMATION AS A SERVICE REV. 1.0

TABLE OF CONTENTS

CONTRIBUTORS

Vijay Ranjan Mungara - Intel Corporation
Bala Rasaratnam - National Australia Bank
Jason, Li-Yi Lin - Taiwan Stock Exchange
Mimosa Tang - Taiwan Stock Exchange
Matt Estes - The Walt Disney Company
Shawn Ramsay - The Walt Disney Company

LEGAL NOTICE

This "**Open Data Center Alliance℠ Usage Model: Data Management for Information as a Service**" document is proprietary to the Open Data Center Alliance (the "**Alliance**") and/or its successors and assigns.

NOTICE TO USERS WHO ARE NOT OPEN DATA CENTER ALLIANCE PARTICIPANTS: Non-Alliance Participants are only granted the right to review, and make reference to or cite this document. Any such references or citations to this document must give the Alliance full attribution and must acknowledge the Alliance's copyright in this document. The proper copyright notice is as follows: "**© 2013 Open Data Center Alliance, Inc. ALL RIGHTS RESERVED.**" Such users are not permitted to revise, alter, modify, make any derivatives of, or otherwise amend this document in any way without the prior express written permission of the Alliance.

NOTICE TO USERS WHO ARE OPEN DATA CENTER ALLIANCE PARTICIPANTS: Use of this document by Alliance Participants is subject to the Alliance's bylaws and its other policies and procedures.

NOTICE TO USERS GENERALLY: Users of this document should not reference any initial or recommended methodology, metric, requirements, criteria, or other content that may be contained in this document or in any other document distributed by the Alliance ("**Initial Models**") in any way that implies the user and/or its products or services are in compliance with, or have undergone any testing or certification to demonstrate compliance with, any of these Initial Models.

The contents of this document are intended for informational purposes only. Any proposals, recommendations or other content contained in this document, including, without limitation, the scope or content of any methodology, metric, requirements, or other criteria disclosed in this document (collectively, "Criteria"), does not constitute an endorsement or recommendation by Alliance of such Criteria and does not mean that the Alliance will in the future develop any certification or compliance or testing programs to verify any future implementation or compliance with any of the Criteria.

LEGAL DISCLAIMER: THIS DOCUMENT AND THE INFORMATION CONTAINED HEREIN IS PROVIDED ON AN "**AS IS**" BASIS. TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, THE ALLIANCE (ALONG WITH THE CONTRIBUTORS TO THIS DOCUMENT) HEREBY DISCLAIM ALL REPRESENTATIONS, WARRANTIES AND/OR COVENANTS, EITHER EXPRESS OR IMPLIED, STATUTORY OR AT COMMON LAW, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE, VALIDITY, AND/OR NONINFRINGEMENT. THE INFORMATION CONTAINED IN THIS DOCUMENT IS FOR INFORMATIONAL PURPOSES ONLY AND THE ALLIANCE MAKES NO REPRESENTATIONS, WARRANTIES AND/OR COVENANTS AS TO THE RESULTS THAT MAY BE OBTAINED FROM THE USE OF, OR RELIANCE ON, ANY INFORMATION SET FORTH IN THIS DOCUMENT, OR AS TO THE ACCURACY OR RELIABILITY OF SUCH INFORMATION. EXCEPT AS OTHERWISE EXPRESSLY SET FORTH HEREIN, NOTHING CONTAINED IN THIS DOCUMENT SHALL BE DEEMED AS GRANTING YOU ANY KIND OF LICENSE IN THE DOCUMENT, OR ANY OF ITS CONTENTS, EITHER EXPRESSLY OR IMPLIEDLY, OR TO ANY INTELLECTUAL PROPERTY OWNED OR CONTROLLED BY THE ALLIANCE, INCLUDING, WITHOUT LIMITATION, ANY TRADEMARKS OF THE ALLIANCE.

TRADEMARKS: OPEN CENTER DATA ALLIANCE℠, ODCA℠, and the OPEN DATA CENTER ALLIANCE logo® are trade names, trademarks, and/or service marks (collectively "**Marks**") owned by Open Data Center Alliance, Inc. and all rights are reserved therein. Unauthorized use is strictly prohibited. This document does not grant any user of this document any rights to use any of the ODCA's Marks. All other service marks, trademarks and trade names reference herein are those of their respective owners.

OPEN DATA CENTER ALLIANCE℠ USAGE MODEL: DATA MANAGEMENT FOR INFORMATION AS A SERVICE REV. 1.0

EXECUTIVE SUMMARY

Information as a service is defined as the ability to provide standardized and secure self-service methods to create, exchange, and extract meaningful information from all available data in the right format at the right time. A critical capability for accomplishing information as a service is the ability to manage and process data; specifically, to source and collect data; to standardize, cleanse, and enrich data to ensure quality and usability; to manage the lifecycle of the data; and to ultimately convert data into information that can provide analytics, predictions, and business intelligence.

In the rapidly evolving data landscape, data management in an information-as-a-service ecosystem encounters several challenges:

- **Big data.** Data is being created and consumed at an increasing pace in today's world. In addition, a variety of new data sources and data formats must be processed, including voice, web, and scientific data; data generated from machines, controls, sensors, and devices (ranging from cars to wearable computers); and structured, unstructured, and poly-structured data.
- **Data location.** Today, data originates not only within the organization and its network(s), but also from the cloud, mobile, devices, and social media. Also, data may originate in a variety of geographical locations. This emerging "boundary-less" computing environment complicates compliance (including privacy rules), security, and data processing logistics.
- **Data time lines.** An ever-increasing business velocity means batch processing is no longer accepted as the norm. Businesses need—and expect—information in real-time or near-real-time, as well as easy access to data, accompanied by easy-to-use, rich analytical capabilities.

This focused usage model offers a way to simplify the complex problem of data management in an information-as-a-service ecosystem by breaking it into four manageable stages: data sourcing and collection, data standardization, data lifecycle management, and information and insight delivery.[1]

This document begins by providing an introduction to information as a service and an overview of data management in an information-as-a-service ecosystem. Then, for each of the four data management stages, it provides a discussion of the relevant challenges, risks, and opportunities as well as a discussion of how processes, people, and tools and technologies can help address these challenges.

This document concludes by enumerating a set of foundational requirements for data management in an information-as-a-service ecosystem and summarizing industry actions that we recommend be taken by solution providers and consumers of those solutions to foster successful data management for information as a service.

[1] These stages were first introduced in Figure 5 in the "Open Data Center Alliance Master Usage Model: Information as a Service." See www.opendatacenteralliance.org/library

DEFINITIONS

Table 1 defines the terms used throughout this document.

Table 1. Terms and definitions.

Term	Definition
CRUD	Create, read, update, and delete settings for data.
Data management	The management of data assets throughout its lifecycle by leveraging architectures and standard policies and practices.
Information as a service	The ability to provide standardized and secure methods to create, manage, exchange, and extract meaningful information from all available data in the right format at the right time.
Master data management	The application of data management practices and tools to define and manage master data (non-transactional entities) within an organization.
Metadata	A set of data that describes data. Metadata is further classified into: • Structural metadata. Describes the structure of the data, including data types and lengths • Descriptive metadata. Describes the content of the data and what it represents. • Operational metadata. Describes the processes that have acted upon the data, including transformations and lineage.
Multitenancy	Multitenancy is the ability to virtually partition a single environment while providing the security controls to ensure isolation of each tenant's data, processes, and workload.
Poly-structured data	Also known as multistructured data, poly-structured data is a form of structured data whose structure may vary from row to row (tuple to tuple) within the course of a query, an update, or execution of a program. Examples of poly-structured data include XML, JavaScript Object Notation (JSON), and Binary JSON (BSON).
Unstructured data	Unstructured data is data that does not fit into a predefined data model. The unstructured data may exist as data in a file or document, or encapsulated within a attribute of a structured or poly-structured data object.

PURPOSE AND AUDIENCE

This focused usage model defines the tasks necessary for successful data management in an information-as-a-service ecosystem. This document serves a variety of groups, including the following:

- Business decision makers who control investment in new processes, human capital, and tools and technologies
- Risk management and security operation teams seeking a means to improve data controls for privacy, confidentiality, loss prevention, and data quality management in complex information architecture ecosystems
- IT groups involved in planning, design, operations, and procurement that want to improve solutions from the perspectives of data quality and data management
- Solution providers and technology suppliers seeking to better understand customer needs and tailor service and product offerings
- Standards organizations involved in defining standards that are open and relevant to end users

SCOPE

This focused usage model discusses data management in an information-as-a-service ecosystem in four stages: data sourcing and collection, data standardization (quality, cleansing, and enrichment), data lifecycle management, and information and insight delivery. Included in the discussion are detailed requirements and the challenges of data management in an information-as-a-service implementation. The discussion of solution considerations involving processes, people, and tools and technologies is at a high level.

The discussion of technologies in this focused usage model is limited to data management. The discussion does not include details of how to transform data into information and derive insights from this information, as this topic is adequately covered in the "Information as a Service Master Usage Model."[2] We discuss the 12 architectural components put forth in the "Information as a Service Master Usage Model" only in regard to how they relate to data management.

[2] See www.opendatacenteralliance.org/library

INTRODUCTION TO INFORMATION AS A SERVICE

The combination of big data (driven in part by mobility and social media) and cloud computing is exponentially expanding the amount and types of available data, while at the same time the growing business demand for near real-time information is compressing the time available to process and use that data. This presents both challenges and opportunities for solution providers and consumers of information. Information as a service, defined as the ability to provide standardized and secure methods to create, manage, exchange, and extract meaningful information from all available data in the right format at the right time, is one of these opportunities.

The benefits of information as a service include the following:

- A dynamic ability to acquire information and access business insights through the orchestration of information delivery from multiple data sources and in multiple formats.
- Reduction of cost, time, and complexity of sharing data stored in multiple locations.
- Standard methods that enable data consumers and producers to interact easily, consistently, and efficiently.
- Separation between the consumers of data (end users, applications, or other services) and the data that they use or produce.

Interestingly, the same forces that are driving the rapidly evolving data landscape are also enablers of information as a service. For example, organizations are using big data technologies in conjunction with existing data warehouse infrastructure to process greater volumes of data on shorter time cycles. These solutions have been fueled further by the greater access to information enabled by mobile solutions. Cloud computing has the most dramatic effect on information as a service, because it has propelled businesses, consumers, and applications to interact with each other across organizational, geographic, and technological boundaries.

The Open Data Center Alliance (ODCA) provided an introduction to and discussion of information as a service in a previous work (see Open Data Center Alliance Master Usage Model: Information as a Service[3]). That document explored information as a service through the lens of 12 architectural components required for information-as-a-service implementation. Information as a service is developing at the crux of several market drivers. Through this evolving model, businesses can more dynamically orchestrate the flow of information across their enterprise and deliver business insights faster, and with less cost.

OVERVIEW OF DATA MANAGEMENT FOR INFORMATION AS A SERVICE

Data management, the focus of this usage model, is an element of each of the 12 information-as-a-service architectural components.[4] Data management is important because it enables standardized, consistent, and controlled access to data. By applying a standard set of transformations to the various sources of data and then enabling applications to access the data using open standards, service requestors can access the data consistently regardless of vendor or system in a secure manner. Instead of building a custom distribution system for each business application, information as a service relies on a conceptual capability architecture that enables integration of back-end data sources and front-end information consumers.

Figure 1 shows a sample architecture that enables efficient data management for information as a service. Each component defines the settings and functions that the administration of a use case can implement and customized. Table 2 lists the definitions of the various components shown in Figure 1.

[3] See www.opendatacenteralliance.org/library

[4] Illustrated in Figure 4 of the "Open Data Center Alliance Master Usage Model: Information as a Service." See www.opendatacenteralliance.org/library

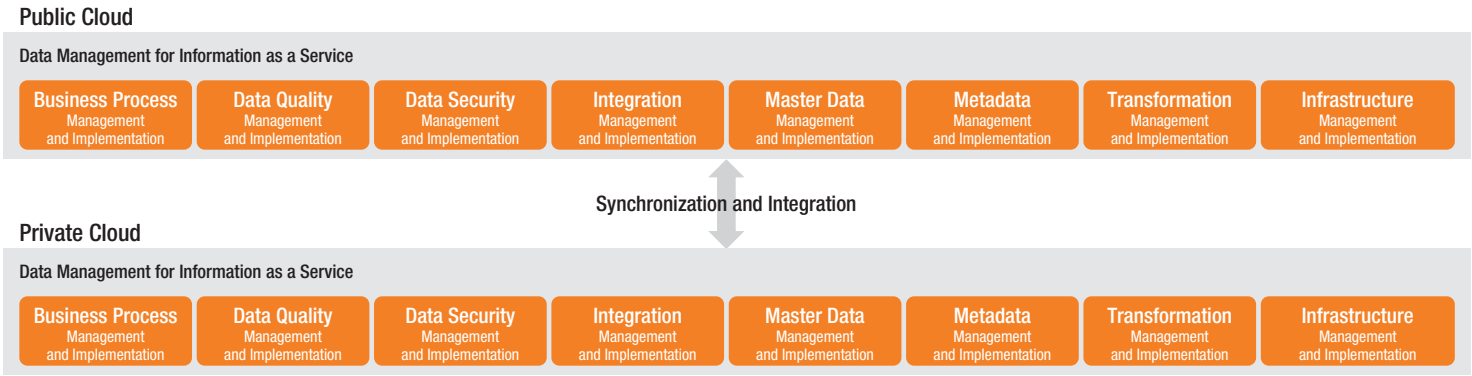

Figure 1. Sample architecture that enables efficient data management for information as a service.

Table 2. Definitions of data management architecture components.

Component	Definition
Business Process Management and Implementation	Allows users to define the create, read, update, and delete (CRUD) functions for data usage. These settings can be customized differently for different use cases.
Data Quality Management and Implementation (see also the discussion of data quality in Table 3 and Table 4)	Consolidates the data quality requirements from all other managers; allows users to further customize and define all data quality requirements (such as allowable NULL values, mandatory values, percent of acceptable data quality issues in the data being used by the use case, and level of required accuracy and precision). This manager can also be used to define the data governance preferences for collection of data (such as whether the data should be cleansed as soon as it is entered and the frequency at which it should be collected (such as once per day or whenever modification occurs).
Data Security Management and Implementation	Protects data from privacy and confidentiality issues, and helps prevent fraud, loss, and corruption through the use of techniques such as data encryption, tokenization, and identity management.
Infrastructure Management and Implementation	Allows users to define acceptable service levels, velocity (real-time or batch processing), backup and recovery, and multi-tenant rules for the use case. These definitions help infrastructure-as-a-service providers define the appropriate hardware requirements for the use case. This approach enables hardware to be easily customized per use case.
Integration Management and Implementation	Allows administrators of different use cases to define data integration needs. For example, this component could define how the data IDs relate between the master data and the transaction data within a process (1-to-many, 1-to-1, or many-to-many).
Master Data Management and Implementation	Allows users to define the master data (including reference data) usage (and possibly CRUD settings) and the data quality requirements for both data sourcing and collecting, enabling a single source of truth and maintenance point for core data.
Metadata Management and Implementation	Manages information about the data and is critical for sustainable data management. It includes business metadata such as definitions, a business glossary, descriptions, derivation rules, and conceptual or logical data models; technical metadata such as physical data models, data structures, and data lineage; and operational metadata such as delivery time, exception and error handling rules, and thresholds.
Transformation Management and Implementation	Allows users to define the transformation requirements for a specific use case—for example, whether read operations occur in XML, NoSQL, or RESTful (REpresentational State Transfer) API format. This component can also be extended to define the structure of certain formats, such as JavaScript Object Notation (JSON), Binary JSON (BSON), and RESTful APIs, for specific data items per use case or per usage scenario. In addition it also enables standardization of data content. For example, customer gender codes can be 1 or 2 in one system and M or F in another system to represent male and female. Transformation can convert all gender codes to M and F to enable consistency.

An appropriate level of data management can reduce the risks associated with data. Data risk results from inadequate or failed internal processes, people, and systems or from external events that negatively affect data quality and data security. According to the Australian Prudential Regulation Authority "Prudential Practice Guide for Managing Data Risk,"[5] the consequences of data risk can be significant and can include the following:

- Impact to business objectives
- Inability of the business to meet financial obligations to stakeholders and customers
- Fraud due to data theft
- Business disruption due to data corruption or unavailability
- Execution delivery failure or incorrect business decisions because of inaccurate data
- Breach of legal or compliance obligations (such as noncompliance with privacy regulations) resulting from disclosure of confidential, or incomplete or inaccurate information

Figure 2 shows the tight coupling of data, information, and operational risk.

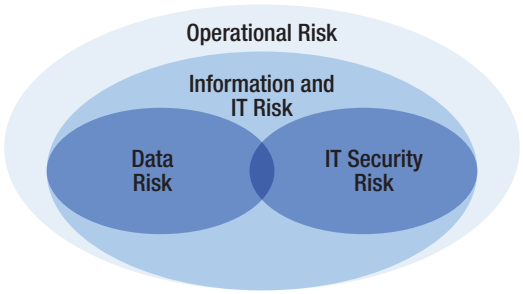

Figure 2. Data risk is a prime contributor to operational risk. © Australian Prudential Regulation Authority, September 2013.[6]

Data Quality Dimensions

Table 3 enumerates several standard elements and principles of data quality. Properly implemented, these can minimize data risk and its impact. Note that the exact level of confidence for each dimension depends on the use case. For example, a marketing campaign may require a different level of accuracy than finance operations or risk management. One way to approach defining an acceptable level of data quality—which includes how much the data can be trusted—is to develop a "scoring" system. For example, high-quality, highly trusted data may score a 5, while lower quality data or data with trust issues may score only a 2.

Table 3. Traditional Data Quality Dimensions that Can Minimize Data Risk.[7]

Data Quality Dimension	Description
Accuracy	The degree of confidence that data is error-free
Completeness	The extent to which data is not missing and is of sufficient breadth and depth for the intended purpose
Consistency	The degree to which common data across different sources follows the same definitions, value ranges, types, and formats
Confidentiality	The level of assurance that only authorized access to data is permitted
Authenticity	The level of assurance that the quality or condition of data is genuine and has not been subject to unauthorized change
Non-Repudiation	The degree of traceability of the data; each data event can be verified, and events cannot later be denied
Timeliness/Currency	The degree to which data is up to date
Availability	The level of assurance that the data is accessible and usable when required
Fitness for Purpose	The degree to which the data is relevant, appropriate for the intended purpose, and meets business specifications

[5] "Prudential Practice Guide CPG 235 – Managing Data Risk" (2013). Australian Prudential Regulation Authority www.apra.gov.au.
[6] Figure 2 is used with permission from the Australian Prudential Regulation Authority "Prudential Practice Guide CPG 235 – Managing Data Risk." www.apra.gov.au
[7] The information in Table 3 is used with permission from the Australian Prudential Regulation Authority "Prudential Practice Guide CPG 235 – Managing Data Risk." www.apra.gov.au

Data Controls

There are various standard data controls that can help to foster data quality, as shown in Table 4. Some of these controls should be set up as part of the technical solution. Others relate to business processes, and some have aspects relevant to both areas.

Table 4. Data controls.

Data Control	Technical Solution	Business Process
Data and information policies, principles, and standards		✓
Overarching governance of policy compliance, including privacy considerations		✓
Data ownership		✓
Data change management	✓	✓
Data quality issue management, such as data quality monitoring, profiling, and cleansing	✓	✓
Reconciliation and error control	✓	✓
Metadata management	✓	✓
Access control	✓	✓
Automation and schedule control	✓	✓
Audit control (traceability audit)	✓	✓

The following emerging industry trends make the information-as-a-service architecture highly complex. Managing data quality and establishing data controls in such a highly complex architecture is extremely challenging. This focused usage model can help organizations surmount these challenges as they build an information-as-a-service system.

- **"Boundary-less" information ecosystem.** When data, information, and insights come from the cloud, third-party service providers, external data, software as a service (SaaS), and other sources, it can be difficult to achieve the necessary accuracy, completeness, ownership accountability, and data quality.
- **Big data.** As new varieties of data sources emerge (such as voice, text, web, and machine logs) and the volume and velocity of data increases exponentially, it becomes necessary to appropriately integrate these sources with the traditional structured data of the organization—enabling an organization to gain maximum value from all available data sources. For example, access to customer sentiments analyzed from call center voice data, linked to customer churn information from traditional customer relationship management systems, can provide more clarity about actions that could improve customer engagement. Establishing a good understanding of the context of data and information (metadata), and achieving consistency between various sources to appropriately integrate the data, is no easy task.
- **Reduced time dimension.** In a complex information-as-a-service ecosystem, a rapidly growing demand for real-time and near-real-time information delivery requires new approaches and technology and tools.
- **Dynamic data activities.** Increased business need for easy access to data in flexible exploratory environments for discovery of new insights and model development add complexity to the management of information security and audit controls. However, such management is extremely important in such environments to minimize data risk while balancing business agility and flexibility.

While data management in a complex information-as-a-service ecosystem is challenging, it is extremely important. The lack of appropriate data management and controls can potentially impact the quality of data, diminishing the value of the information and insights delivered to end users. In turn, that diminished value can negatively impact end-users' decisions and actions, and can significantly increase the organization's exposure to operational, regulatory compliance (such as breach of privacy requirements), and reputation risks.

Figure 3 shows a sample data framework that helps establish appropriate data controls in an organization to manage data and minimize data risk.

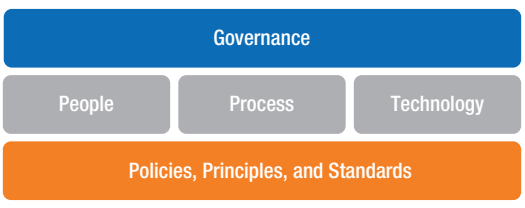

Figure 3. This sample data framework can help establish appropriate data controls.

DATA MANAGEMENT STAGES

Each data management stage—sourcing and collection, standardization, lifecycle management and insight delivery—is presented separately in this document, and the traditional data warehouse and business intelligence (BI) data flow shown in the Open Data Center Alliance Master Usage Model: Information as a Service[8] and depicted at the top of Figure 4 represents a linear flow. However, the stages are more intertwined and not necessarily performed according to the traditional data warehousing processing patterns, tools, and techniques. While traditional processing architectures provide a great deal of value, the ability to inexpensively store and process large quantities of data provides an opportunity to rethink the traditional data flow. While all the same processes still exist, the added capabilities allow the business to put them together in different orders.

Figure 4 shows several data flows that support the new patterns of data supply chain and new industry trends (other data flows are possible). In the figure, the "late binding" flow includes processing and standardization *after* storage, as part of delivery. This data flow, supported by increased processing capabilities, leaves data in its native unstructured or poly-structured format—applying standardization techniques when the data is queried for presentation in the information delivery layer. This approach allows multiple business processes to access the data in its native format and use the data flow and transformation techniques that make the most sense for a particular business process—thereby enabling greater reuse of the data.

For the streaming flow illustrated in Figure 4, there is no data storage at all. Instead, data flows directly from the source (quite often, machine-generated data) to the transformation and information delivery stages.

The subsequent sections of this document describe the various stages of end-to-end data management, including the challenges associated with each stage as well as solution considerations associated with people, processes, and tools and technologies that can help minimize data risks in an information-as-a-service ecosystem.

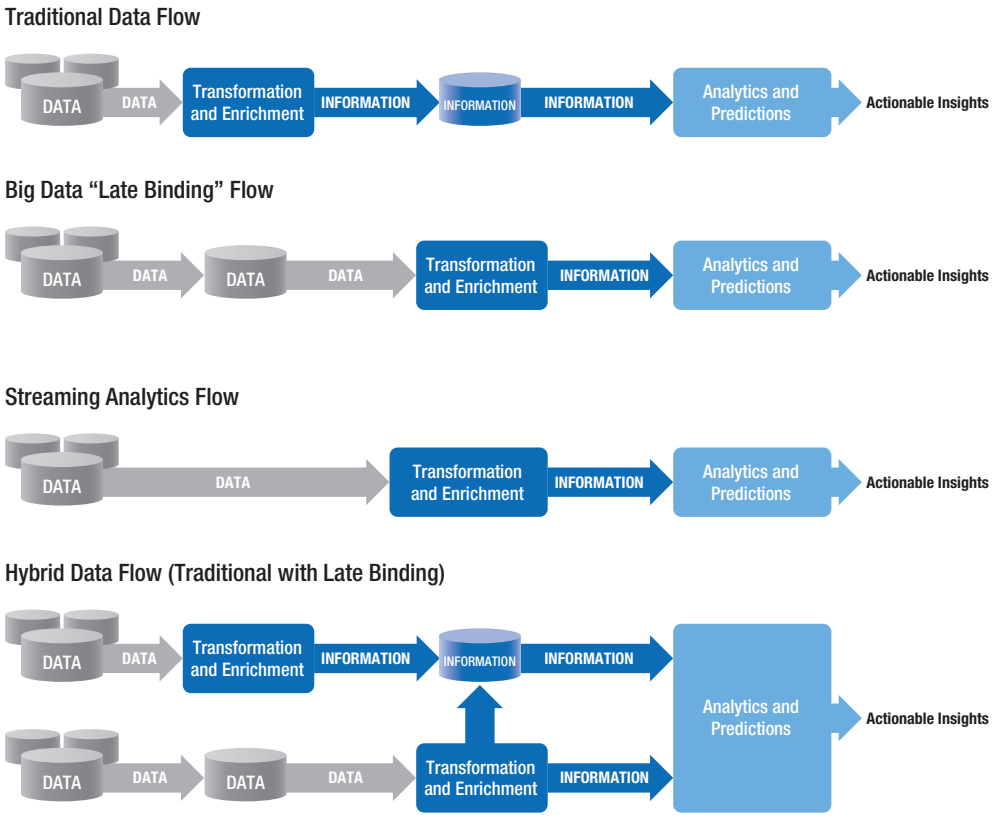

Figure 4. Several data flows are possible, depending on the data and use case.

[8] Shown in Figure 5 in the Open Data Center Alliance Master Usage Model: Information as a Service. See www.opendatacenteralliance.org/library

Data Sourcing and Collection

Data sourcing (the process of determining the appropriate source for needed data) and data collection (the process of actually accessing the data once the source has been identified) present challenges in an information-as-a-service ecosystem, due to the boundary-less location, variety, velocity, and volume of data inherent in such a system. Information as a service by its very nature implies a hybrid information universe, with multiple sources, formats, data owners, and service levels. In this hybrid environment, some data comes from external sources such as a public cloud or software-as-a-service providers. Other data comes from within the organization.

Challenges and Opportunities

Business processes, tools and technologies, and support systems raise common challenges to both data sourcing and data collection. Because information-as-a-service ecosystems potentially involve several entities external to a particular company, the roles performing the business processes are not limited to individual people but instead represent entire departments and/or organizations, which in turn increase the complexity in dealing with service availability, licensing, and costs. The complexity also affects obtaining business process metrics, such as time-to-process completion, process-step quality, and process I/O quality, and in meeting data security and privacy requirements within each use case.

Existing tools and technologies being used in the information-as-a-service ecosystem must evolve to use and collect multiple data formats and support multi-tenant use cases. In some cases, an organization may need to purchase or develop new tools; in other cases, updating existing tools may be a better choice.

Support systems for data sourcing and collection issues and incidents within an information-as-a-service ecosystem must address data quality, data management, and governance. These aspects must be integrated into the overall business processess, and ideally support should be centralized even though the data may come from multiple sources. Other questions that focus on data quality and governance include the following:

- Should data quality governance be automated and/or outsourced?
- Should privacy and security restrictions be applied based on use cases?
- In the case of big data platforms, should the restrictions be applied while collecting the raw data or after it is collected?
- If there are variations in restrictions depending on use cases, how can compliance be verified?
- How can the system enable flexibility, whereby restrictions can be customized based on the use case?
- What is the right balance between data sourcing frequencies and data standardization expectations?

As business processes, tools and technologies, and support systems evolve, organizations need to develop a strategy for transition change management.

Other common challenges to data sourcing and collection revolve around how the sourcing and collection solutions are architected:

- **Reusability.** Organizations should consider whether to make the data single-tenant or multi-tenant. The answer depends on the use case, but the nature of information-as-a-service ecosystems predisposes the assumption that the systems need to work anywhere in the world (any premise). Sourcing data from a public cloud information-as-a-service system raises significant challenges in providing the ability to integrate with data and business processes within an organization's private cloud or information-as-a-service system. In many cases, business processes define the boundary of the information being used, and depending on the use case, business processes often differ. Organizations need to consider how to make information provided as a service generically applicable to most business processes.
- **Data format.** Related to the prior bullet, data format is also a challenge for data sourcing and collection. Because many systems get source data from multiple sources as well as in multiple formats, it can be difficult to convert the raw data into information that can be consumed generically across several use cases. For example, sourcing data in a relational database management system (RDBMS) format may not be useful to real-time BI use cases. On the other hand, sourcing unstructured data using the NoSQL or NewSQL formats may not be useful to transaction-based use cases. However, converting unstructured data to a structured format and integrating it with structured data can provide extremely rich customer information and insights.
- **Standardization.** Deciding exactly when standardization and enrichment should occur depends on the use case. Error control at the capture point is required for some use cases, but cleansing isn't always necessary for other use cases. In particular, emerging real-time, near-real-time, and change data capture integration capabilities pose significant challenges related to data standardization and cleansing—the more standardization, the less timely the data. For more details, refer to Figure 4 and the discussion of Data Standardization in a subsequent section.

- **Bandwidth requirements.** Sourcing and/or collecting data downstream raises bandwidth (volume and velocity) issues. The question becomes whether to send all data at once or to send some data in chunks. Either way, it is important to minimize the number of necessary extracts. Also, organizations need to decide whether using in-memory databases is the best solution for bandwidth issues or whether it is better to use more traditional solutions to temporarily store data locally instead of in the cloud. The answer to these types of questions partially depends on the use case's latency requirements. Some use cases may need real-time or near-real-time responses, while other use cases may be able to accommodate more lag time. When deciding on data management approaches for information as a service, organizations must weigh cost against requirements and enable an elastic infrastructure that can dynamically adjust to each use case's needs.
- **Access to metrics.** Given the external and internal integration involved in many information-as-a-service ecosystems, organizations also may need to plan for how they will access process metrics. These metrics track process characteristics such as how much time it takes to obtain data, who obtained the data, tool performance, and the quality of data while performing a particular process step.

Solution Considerations

To address the above challenges, the information-as-a-service data management architecture needs to accommodate components that manage various stages, as shown earlier in Figure 1. These components define, on a use-case basis, the acceptable parameters, preferences, and settings for the information-as-a-service ecosystem. For example, the Business Process component could define CRUD settings for data depending on how it is being used.

Process: The data management architecture for an information-as-a-service ecosystem should include the ability to customize the following aspects of data sourcing and collection, per use case.

- Read operations and collection times
- Definitions of roles, organizations, and decision points, including data governance roles
- Definitions of the process metrics that are of interest
- Acceptable service levels

Also, sourcing and collection solutions should include the following:

- **Data controls in an end-to-end data flow.** Organizations should embed appropriate data controls in their information delivery processes (see Table 4). Examples include data reconciliation between appropriate source and target systems, file-level controls (such as verifying that the file with the correct day and time stamp is processed, or performing record-count validation for completeness), and establishing appropriate schedule control to ensure the data flows end-to-end with the right sequence of automated processes that deliver to the business the right information at the right time.
- **Error thresholds and management.** Organizations should establish data quality thresholds and alerts for when these thresholds are reached across the information-as-a-service ecosystem. At a minimum, these should be established wherever business-critical data is stored.
- **Data as a service.** Acquisition and distribution of a wide variety and volume of data (with history) can be offered as a dedicated service to the end-user community by internal and/or external service providers.

People: The following roles are useful in establishing and maintaining proper data management in an information-as-a-service ecosystem:

- **Data subject matter experts (SMEs)** have an understanding of the context of data and its business value.
- **Data Integration specialists** can design appropriate solution patterns for various integration use cases, based on business expectations. Such patterns include real-time, near-real-time, and batch processing.

Tools and technologies: Organizations should choose technologies that can easily scale up or down, depending on the use case. Also, the following components provide significant value to data sourcing and collection in an information-as-a-service ecosystem:

- **Enterprise data model.** Defines all business-critical data and provides information about the data owners and quality expectations for specific data domains and entities. For example, the customer data domain may require 100-percent accuracy of the customer address, but may not expect as much accuracy in customer tweets.
- **Data catalog.** Stores authoritative sources and data service providers for each of the various data entities and types.

- **Use case-specific tools.** Different tools and technologies excel in different use cases. Depending on an organization's needs, their data sourcing and collection solutions should include tools that support extract-transform-load (ETL) and extract-load-transform operations, change data capture, and real-time service buses.

In addition, the following interfaces should be integrated into the information-as-a-service data management solution so that requirements, definitions, and processes can be customized on-the-fly for a particular use case.

- Support for customization of performance and multi-tenant/single-tenant requirements
- Ability to define sourcing and collection data format configurations
- Ability to define standardization and governance requirements
- Ability to define privacy and security considerations
- Ability to dynamically implement within the information-as-a-service infrastructure a particular use case's process, roles, and tool customizations and configurations during the execution of the use case.

Data Standardization

The key elements of data standardization include traditional ETL processing, standardization techniques, and augmentation techniques. These are described further in Table 5.

Table 5. Key elements of data standardization.

Element	Examples
Traditional extract-transform-load processing	• Selections of data to process. • Translating coded values. • Encoding data. • Derivation of calculated fields. • Lookup and validation. • Data type conversion.
Data standardization techniques	• Pulling apart data and standardizing formats. – Names – Mailing addresses – Phone numbers • Transformation of freeform fields. For example, taking a field with "M," "F," "Male," and "Female" and converting it to the standard ISO/IEC 5218 format. • Data format standardization. For example, data reduction using tools such as MapReduce, converting unstructured data to a structured format, and converting voice data to text.
Data augmentation techniques	• Pulling data from another source based on the data being processed. For example, pulling in geo-demographic information based on a standardized mailing address. • Updating or adding data based on heuristics. • Tagging common records based on a fixed set of rules. • Integrating reference data, such as hierarchies and classification data. • Using data mining models to calculate a score on the probability of events happening. • Using natural language processing and sentiment analysis to pull data from unstructured data elements.

Challenges and Opportunities

With information as a service, a wide variety of internal and external data must be standardized and integrated to gain maximum value from the data. But accomplishing this in huge volumes and at a staggering velocity of creation, capture, and change poses a significant challenge for data management. When considering data standardization in an information-as-a-service ecosystem, some of the challenges include the following:

- How does an organization determine when and where to perform standardization? In other words, what data flow (see Figure 4) makes the most sense for a particular grouping of data and a particular use case? In some cases, it may make sense to transform the data into value-added information right away, or to source the data raw in its native structures in simple consumable formats such as XML, JavaScript, Object Notation (JSON), or delimited text, and then standardize it later. A rapidly increasing need for near-real-time information delivery demands integration of data closer to the point of delivery, hence the "late binding" and hybrid data flows shown in Figure 4.
- If standardization is performed later rather than sooner, how does that affect data integration? For example, it can be difficult to integrate data representing different levels of granularity, such as individual customer-level data with aggregated and summarized data at a customer-segment level. Or, it can be challenging to integrate data items that are generated at different points in time.
- How does an organization ensure that value-added information is valued commonly across all use cases? Value-add can mean different things to different consumer systems. For example, some use cases use tangible value-added techniques such as measuring return on investment in terms of cost savings; others use intangible value-added techniques such as reduced time-to-process or process optimization.
- How does an organization deal with trust issues that arise during standardization, such as when the data format is changed?
- How can the organization utilize data from external systems and services for use in the data standardization and augmentation processes?
- How can the organization measure data standardization in a distributed processing environment and use that information to improve data management and data transformation processes?

Solution Considerations

To answer these types of questions, an organization can evaluate processes, people, and tools and technology to find solutions.

Process: Where standardization occurs in the data process is dependent on the pattern that best solves the business problems an organization is trying to address. There has been much discussion about which approaches work best for certain problems. Those discussions won't be repeated here; for a sampling, readers can refer to the links listed in the Further Reading section. In general, the industry seems to be acknowledging that a large enterprise will likely leverage multiple patterns and that IT needs to address the issue of presenting the data to the user in a cohesive fashion.

People: It may be necessary to provide training to people working with data—both at the source level and at the consumption level—to educate them about the possible data flows and what level of standardization is possible or beneficial.

Tools and technology: An organization should consider the following concepts when formulating a data standardization plan.

To support streaming processes for in-stream and "late binding" processes, data standardization tools need to provide the flexibility to be executed as part of any portion of the information lifecycle.

- To support the parsing of information from unstructured and poly-structured data sources, standardization tools need to support more complex data manipulation algorithms, such as natural language processing and sentiment analysis.
- With the influx of unstructured and poly-structured data, standardization processes need to be extremely robust and handle errors in a flexible manner, allowing the "good" data to make it through while providing mechanisms to address and reprocess the "bad" data.
- Because the standardization process may be incorporated in various places within the data flow, the ability to share metadata becomes more critical. The tools need to provide clear process metadata around the standardization of the data, including structural metadata about the transformations made and where they occur within the information lifecycle.
- Because the data standardization processes can be directly embedded in the analytical processes, the tools may get paired with data profiling and exploration tools. This makes it critical that the standardization and profiling tools be able to easily share and integrate their metadata, allowing for initial data analysis to be incorporated into the processing rules in the data standardization tools.
- The distributed nature of cloud processing provides for API access to a wide variety of services, including those that are useful in data standardization. The standardization tools need to support calls to these external services as part of their workflows. Likewise, these tools should expose their functionality to external tools using standard services.

Data Lifecycle Management

Data lifecycle management (DLM) is a series of data management activities that form a comprehensive approach to manage an organization's data throughout its lifecycle. DLM involves procedures and practices as well as applications. Key elements of DLM include the following:

- **Data classification.** When data is created or collected, it has to be classified into different categories in accordance with the aspects of data, such as data sensitivity and risk.
- **Data catalog.** DLM must maintain a list of unique data items.
- **Data rights (or access control) management.** This includes management of data ownership and access control, such as data sharing, revocation of access rights, and sharing period.
- **Data lifetime and relationship management.** DLM must deal with two issues after data are aggregated, processed, and stored: the management of data lifetime and the relationship between different data items. The owner of the data needs to determine the lifetime of data. The relationship between different data has to be built into a diagram (or record) that should be updated whenever any change of relationship is made.
- **Data preservation.** This includes backup and restore procedures, and synchronization of data between different data centers.
- **Content discovery.** When data are aggregated and stored, a method or tool should be provided for applications or users to find the data they want to use. This tool or method should leverage both metadata and the business glossary.
- **Data deletion.** When the lifetime of the data is reached, DLM should ensure that the deletion of the data will not be a problem to other related data—this should occur before the deletion. Once the deletion is ready to occur, DLM should ensure that all instances of the data is deleted (such as data that resides in several data centers).
- **Data security.** In addition to the data control measures already mentioned, data security also includes data encryption at rest and in transit, data integrity, and data availability.
- **Auditing.** This includes data usage monitoring and logging to meet all related legal and compliance requirements, such as verifying that proper privacy controls are in place.

Challenges and Opportunities

Information-as-a-service ecosystems present several DLM challenges and opportunities, because of the diverse sources and formats of data:

- **Big data.** While the big data phenomenon is giving organizations a broad range of new options for data analysis, it also compounds challenges associated with managing, organizing, and protecting data. Enterprises face technical challenges in managing rapidly expanding volumes and various types of data.
- **Cloud computing.** Cloud computing brings new issues to enterprises, such as the following:
  - The multi-tenancy issue involves the problem of how to isolate data stored in the same cloud environment by different sensitive levels of data and different users, and how to assure that the deletion of data is performed completely and will not be recoverable.
  - Since data may be stored in different locations for different purposes (for example, load balance and backup), the consistency of data must be assured when it is being updated, deleted, and synchronized so that applications or users will not misuse the outdated data to generate inaccurate information. Inaccurate information may cause risks to the enterprise.
- **Complexity of the data relationship.** Since the data volume and sources might grow larger as time goes on, the data relationship will become much more complex and difficult to maintain. Meanwhile, one change of a single data source may affect a large amount of other related data.
- **Data consolidation.** It is difficult to consolidate data if an enterprise doesn't have a clear picture of the list of data items and data relationship information. Without a list of data items, there might be a redundancy of the same data item thus raising the possibility of data inconsistency.
- **Data security.** Data security is a major aspect of the information-as-a-service ecosystem, and it is the subject of several ODCA documents and usage models. Refer to the ODCA website[9] for a list of relevant documents.

[9] See www.opendatacenteralliance.org/ourwork/usagemodels

Solution Considerations

The following processes, people, and tools and technology can be considered:

Process: Organizations should consider developing an enterprise data model, which combined with data modeling techniques and methodologies, helps store data in a standard, consistent, and predictable manner—fostering the ability of organizations to manage data as an asset. (Refer to the Open Data Center Alliance Master Usage Model: Information as a Service[10] document for more information on enterprise data models.)

In addition, organizations should establish a data change management system that can track every change in the characteristics of all data items. The characteristics could include the lifetime, owner, and classification of the data. In addition, the system should be able to summarize the relationship between all the data items. When there are plans to remove data that is related to other data items, the system should show a warning message to data owners and data catalog managers. Well-defined corporate policies can help to foster appropriate data preservation and access over prolonged periods.

People: The following play a significant role in DLM for an information-as-a-service ecosystem:

- **Chief data officer.** The chief data officer has a significant measure of business responsibility for determining what kinds of information the enterprise will choose to capture, retain, and use, and for what purposes.
- **Data catalog managers.** The data catalog managers are responsible for maintaining the data catalog as well as the data relationship whenever any change is made (for example, defining new data sources, changing the lifetime of data, or authorization changes).
- **Data owner(s).** Every data source (or item) must have a data owner to define standard business rules for data calculation, derivation, and enrichment and manage the access control list, lifetime, and other characteristics of data. They should establish error tolerance levels and thresholds appropriate for specific types of data, based on business usage and criticality (for example, the data quality expectation for financial or risk regulatory data needs to be 100-percent accurate, whereas the data quality for social media data can be less accurate). When a new data item is generated from other existing data sources, a data owner should also be assigned to this new data item.
- **Data curator.** A data curator can facilitate the use of data and must know the following: what data is available; where data is located; who owns any intellectual property associated with the data; what the security level and control of the data is; and how the data relates to other available data.
- **Data stewards** provide change management for any data changes from the data source(s) and mitigate the downstream—and upstream—impacts.

Tools and technology: Organizations should investigate tools that can support data loss protection procedures, data redundancy, and referential integrity management. Tools should also be in a place that help to promote data privacy and confidentiality. Examples of such tools include the following:

- **Data privacy protection.** Organizations need to adopt a policy-driven, on-demand transformation approach to proactively protect data privacy and support compliance—especially in this new era of "boundary-less" computing.
- **Data relationship analysis.** This type of analysis provides the capabilities to analyze data relationships across applications, discover all or specific data relationships, identify hard-to-find relationships defined and enforced by the application logic, and promote consistent data administrative activities.
- **In-memory database.** An in-memory database primarily relies on main memory for computer data storage. Main memory databases are faster than disk-optimized databases because the internal optimization algorithms are simpler and execute fewer CPU instructions. Accessing data in-memory eliminates seek time when querying the data, which provides faster and more predictable performance than disk-based data access.
- **Storage resource management (SRM).** SRM involves optimizing the efficiency and speed with which a storage area network (SAN) utilizes available drive space. SRM identifies underutilized capacity, identifies old or non-critical data that could be moved to less expensive storage, and helps predict future capacity requirements.

[10] See www.opendatacenteralliance.org/library

Information and Insight Delivery

The key elements of information and insight delivery include the following abilities:

- Perform a variety of business analytics to extract business insights. Examples of analytic techniques include what-if analysis; correlation between events, activities, geographic locations, and outcomes; patterns and prediction analytics; sentiment analysis; slicing and dicing information in multiple dimensions; drill-down analytics; anomaly detection; and content search.
- Visualize and publish information and insights in a simple and quick summary view. Techniques that support this ability include dashboards, graphs, cubes and slicing and dicing, and maps.

Challenges and Opportunities

The existing challenges of delivering information and insight are compounded by the increasing volume and variety of data, new technologies that have recently emerged, and a growing expectation by users to be able to access data in real time or near-real time. While these forces pose exciting opportunities for information and insight delivery, they also pose significant risk to data management for information as a service.

Emerging analytics: Big data analytic capabilities create a great opportunity for the organization to access extremely rich customer and business information on a large scale and in a cost-efficient manner. These new analytic capabilities can lead to a better understanding of customer and business needs, and can enable appropriate and timely decisions and actions. Here are just two examples:

- Customer transaction analysis (provides a view of customer spending patterns and the products they purchase) combined with click analysis from a customer's Internet activities provide more timely marketing opportunities.
- A company can use social media to understand the type of products customers prefer and combine this information with an analysis of the organization's product profitability to drive innovation and create products that meet customer expectations as they evolve over time.

In addition to traditional reporting and analytics in production systems, business demand is rapidly increasing for "undiscovered and non-repetitive" analytics. These include capabilities such as the following:

- Data discovery (discovering and innovating new insights from new combination of data)
- Quick ad hoc or one-time-only insights (determining the success of a specific marketing campaign)
- Prototype solutions
- Development and validation of propensity and predictive models (modeling risks or marketing approaches)

Most of these techniques are considered "non-production" or "sandbox" approaches to gaining insight into data. All of these capabilities require agile delivery of a broad range of data, with business flexibility to "trial" outcomes and derive new insights quickly. Having the ability to effectively integrate and analyze a broad variety of data enables organizations to achieve a "360-degree view" of customers and their needs within a very short time, which can represent a significant competitive advantage.

Data presentation: The advent of mobile applications is creating a new channel for data presentation. As the number of mobile applications—and mobile devices—increases, the demand for more insights to be published through these applications and on these devices is also increasing. As businesses build an information-as-a-service ecosystem, they should consider how they can cater to new mobility trends and a new generation of customers that demand information and insight "on the go." A highly complex information-as-a-service ecosystem can take advantage of SaaS, third-party data suppliers, and cloud computing for storage, processing, and delivery of information and insights in a cost-efficient manner. Having a consolidated and consistent presentation data layer, along with common APIs, digital platforms, and visualization tools, can help an organization maximize the value of data. The valuable insights gained from the data can increase revenue, improve customer experience, increase the organization's productivity, and reduce risks.

Data risk: Data quality issues at any point in the data supply chain can impact the value of the information and insight delivered. For example, in an information-as-a-service ecosystem, data can be generated in the cloud, and by software packages, external service providers, and internal systems. This complexity can lead to a lack of understanding related to data context, insufficient data accuracy, consistency and completeness, and insufficient access. Data ownership and accountability, change governance, service-level agreement (SLA) management, and data quality incident management can also be at risk. In the traditional data warehouse paradigm, data ownership determinations were limited to team boundaries. For example, if a data problem was identified, the discussion might have focused on whether the problem belonged to the data warehouse team or the source application team. But data problems can be challenging to resolve in an information-as-a-service ecosystem, because data ownership could involve several companies in widely varying geographies.

Balancing business agility and flexibility with data risk management and control can also raise issues. An organization must consider access control and privacy and confidentiality management, providing enough access to data to enable insight delivery, but maintaining enough control to protect that data. Also, although "undiscovered and non-repetitive" analytics in an exploratory (non-production) environment can provide significant business value, these analytics should not transform into mission-critical repetitive services without the existence of business-managed controls that provide for disaster recovery; IT controls for security and auditability; incident management; and end-to-end SLAs. Further, acquiring and publishing data to and from various mobile applications can also expose the organization to data security threats. Appropriate data policies, architecture and security standards, controls, and governance processes must be established to minimize these data risks.

Solution Considerations

By adjusting processes, investing in human capital and training, and exploring new tools and technologies, an organization can take advantage of the opportunities associated with information and insight delivery in an information-as-a-service ecosystem, while at the same time minimizing data risk.

Establishment of the following processes is recommended:

- **Data discovery and exploration process.** The organization should establish an agile process for trialing new actionable insights quickly from available data (internal and external). The process should include the appropriate level of governance of how the data is provisioned for discovery activity, with appropriate data security measures (such as data masking) where applicable. The process should also include post-activity closed-loop analysis of the discovery outcome and a process for safely deleting or disposing of the data used for the given discovery activity at completion. If the analysis indicates the activity is valuable, it should include provisions to transition the initiative from the exploratory environment to a production system—making it part of business-critical reporting.
- **Information delivery service management.** In an information-as-a-service ecosystem, it is critical to establish an appropriate end-to-end service operating model with clear service-level expectations. Agreement on the quality and timeliness of data delivery is fundamental. Having the appropriate metrics for service quality, executing quarterly reviews to assess service performance against expectations, and having clear rewards and implication processes to manage service providers based on performance metrics will help reduce potential data risks such as data accuracy, completeness, timeliness, and availability of data in a sustainable way.
- **Data controls in an end-to-end data flow.** Organizations should embed appropriate data controls (see Table 4) in their information delivery processes. Examples include data reconciliation between appropriate source and target systems, file-level controls (such as verifying that the file with the correct day and time stamp is processed, or performing record-count validation for completeness), and end-to-end schedule controls (ensuring that processing steps are performed in the right order and none are skipped). In addition, if the data is presented, visualized, and published to digital platforms and mobile applications, it is critical to establish secure networks, appropriate authentication, authorization, and audit controls to avoid data security threats. These data controls will help to promote accurate, consistent, and controlled delivery of business-critical information and insights.
- **Error thresholds and management.** Organizations should establish data quality thresholds and alerts for when these thresholds are reached across the information-as-a-service ecosystem. At the very minimum, these should be established wherever business-critical data is stored.
- **Data quality monitoring.** Data quality processes should include clearly defined business-critical data with appropriate periodic data profiling, as well as cleansing activities in systems that hold business-critical data.
- **Data change management.** An organization's information ecosystem changes when new data is introduced, when the content or format of the existing data is changed, or when a data feed or content is removed. Such changes can have a major impact on the delivery of upstream and downstream processes and related data feeds. For example when an organization decides to move their call center function to an external service provider, the data feed from existing internal call center applications will be replaced by the new service provider's data feed. The new application's file structure, data naming convention, and data content values may be different from those associated with the original internal application. Appropriate impact assessment needs to be done across the information ecosystem and across all affected processes in the end-to-end data flow, to analyze whether the processes are appropriately amended to accept the new data feed. These processes can include data integration, enrichment, and information and insight delivery.
- **Data quality incident management and remediation.** In a complex information ecosystem, the data consumption and information delivery point could be far from where the data originates. For example, a front-line customer officer might capture the customer address data in the source system. This data is then integrated with additional customer details such as customer needs and is then delivered to the marketing department for a customer sales opportunity. Data quality issues in a highly complex information-as-a-service ecosystem can potentially have

a "domino effect" throughout the information system. Organizations should develop a well-defined process to manage and remediate such data quality issues across the entire data architecture, no matter how complex.

- **Security monitoring.** These process should include real-time monitoring of data and information security threats (such as fraud, identity theft, and malware), as well as alerts for data privacy and confidentiality breaches.
- **Internal and external audits.** Organizations should assess the effectiveness of data controls and data management processes.

Having the right people with the right training is critical to the proper delivery of information and insight in an information-as-a-service ecosystem. Here are some examples of useful roles and training:

- **Data subject matter experts (SMEs)** have an understanding of the context of data and its business value.
- **Business SMEs** can provide the appropriate business context, guiding delivery of required insights and information.
- **Data scientists** can derive meaningful and fit-for-purpose insights by appropriately correlating data, building predictive patterns, and extracting sentiments. These team members can help increase the value of the underlying data because even if the underlying data is accurate, the value of insights can be impacted by how they are derived.
- **Data risk management staff** understand the data risks and regulatory and compliance expectations (such as helping to protect privacy) and can provide guidance and governance for information delivery solutions.
- **Information security specialists** with strong capabilities relating to the concepts, techniques, models, and tools can help minimize information security risks.
- **Operational data quality staff** should have skills in data quality monitoring, profiling, cleansing, analysis, and tools and techniques.

As organizations build their information and insight delivery solutions, the following tools and technologies may be useful:

- **Business glossary.** Stores common definitions and business rules for business data elements. (Examples include customer segments and a calculation of customer and product profitability.)
- **Master data management and reference data management tools.** Enable centralized management and maintenance of core data as well as decentralized or federated use of core dimensional data. (Examples include customer names; products; customer contracts, service arrangements, and orders; product hierarchies; and organizational hierarchies.)
- **Data lineage and technical metadata tools.** Show the end-to-end flow of data—from the point of acquisition through integration, transformation, distribution, and eventual presentation or end-user access. These tools assist data change impact analysis and data quality incident management.
- **Data quality tools.** Provide the ability to profile data against business-defined quality rules, deliver data quality scores, and dashboards for the business to monitor the quality of business-critical data, take appropriate actions, and—when required—cleanse poor quality data.
- **Service management tools.** Enable monitoring of information and insight delivery times and compare them to the relevant SLA. If incidents or failures occur, these tools enable workflow triggers to notify the key stakeholders in the data supply chain so they can analyze the break point and remediate it appropriately.
- **Data visualization, reporting, and presentation.** Enable the publishing of data in a secure way (including access controls) and link to the business glossary to provide end users with the definitions and context for the delivered insights.
- **Data analytics tools.** Enable effective analytics to integrate and analyze structured, unstructured, and poly-structured data appropriately and deliver insights with built-in rules that reduce the risk of human errors. Examples of such tools include data correlation, advanced analytics and model development, and sentiment analysis.

RFP REQUIREMENTS

The ODCA brings consumers and providers of data center technologies together to help advance capabilities in the market. These groups work together to help address challenges in today's solutions and outline a future state for the industry to work toward. One technique utilized in defining this future state is the compilation of request for proposal (RFP) and request for information requirements. Solution providers can use these requirements as input into their product roadmaps. Companies that want to acquire some or all of the components of their information-as-a-service ecosystem can use these requirements as-is, reducing the time required in the procurement stage.

Data management for information as a service involves multiple functional capabilities. Each of these capabilities has an associated set of RFP requirements. Rather than detailing all of these requirements in this document, the ODCA Data Services Working Group has outlined key requirements through a general description. Specific, individual RFP statements can be found through the ODCA Proposal Engine Assistant Tool (PEAT).[11]

First and foremost, it is recommended that all solutions relating to information as a service include the ODCA principle requirement: the service is open and standards-based. It is also recommended that all solutions provide a strong set of security capabilities in the spirit of the ODCA Security Working Group Usage Models.

In general, data management for information as a service should include the ability to manage the data within and passing through an information-as-a-service ecosystem. Each of the following requirements and abilities should be customizable, depending on the business process or use case.

- Authenticate and retrieve data from a variety of data sources, within the boundaries of the data center, external to the data center, and across a number of cloud and third-party providers
- Transform and integrate data to and from a variety of formats, technologies, and use cases
- Profile and understand the nature and quality of individual data items and collections of data (such as files, tables, documents, blobs, and images)
- Assess the accuracy, completeness, and consistency of data and take appropriate action
- Create, integrate, and manage master data across the information-as-a-service ecosystem
- Standardize and augment data at various stages in the data management flow
- Support DLM, addressing the challenges of a heterogeneous technology environment (that might include big data, relational database systems, NoSQL, and more), policy enforcement, and support for key DLM roles (including data owners, curators, and catalog managers)
- Support the creation of insights from raw data, and provide delivery and distribution capabilities
- Support a strict security model with particular attention to regulatory and compliance requirements (such as privacy regulations)

Today's world presents numerous challenges to managing data. The scale and pace at which data is being created, the variety of data formats, and the number of data technologies complicates an already challenging situation. Yet organizations must tackle these challenges or face being outpaced by their competitors or negatively affected by regulatory or other external factors. Organizations require solutions to manage data in the same manner as they manage other important corporate assets. They require a data management strategy, tools, technologies, skilled staff, and proper leadership for data management to be executed efficiently and effectively.

Solutions that address data management for information as a service as described in this usage model are the types of solutions that organizations will turn to. With this motivation in mind, the ODCA encourages solution providers to implement these capabilities within their products and services, and for consumers of those solutions to request the RFP requirements summarized here and detailed in the ODCA PEAT tool.

[11] www.opendatacenteralliance.org/ourwork/proposalengineassistant

SUMMARY OF INDUSTRY ACTIONS REQUIRED

This document takes a deeper look into data management for an information-as-a-service ecosystem and provides a summary of requirements for the solution provider community to consider when delivering data management for information-as-a-service solutions. It also provides strong guidance to the consumer organizations that deploy data solutions—the cloud subscribers themselves. The ODCA Data Services Working Group provides this document as a means of moving toward the goal of fostering collaboration between information-as-a-service solution providers and large enterprise consumers of information services. In the interest of accomplishing the mission of the ODCA, we hope to motivate these two groups to work together to define the open specifications, formal or de facto standards, and common intellectual-property-free solution designs, all of which are required if we are to realize an open, interoperable, and thriving market of information-driven businesses and information-driven services.

The following actions are required by the combined solution provider and consumer communities:

- Development and adoption of open and intellectual-property-free reference architectures for data management for information as a service, with particular emphasis on orchestration and automation of the data management flows presented earlier in this document.
- Investment by solution providers in delivering solutions that meet the requirements of this and other usage models as outlined by the enterprise consumers participating in the Open Data Center Alliance.
- Continued feedback from the consumer community on the features and capabilities that are most important to the success of their business, and a continued consolidation and integration of capabilities by solution providers of big data, NoSQL, NewSQL, RDBMS, data warehousing, BI, analytics, metadata, master data, ETL, data virtualization, and other data-related technologies.
- Use of ODCA requirements by solution providers as they develop their products and solutions, as well as a commitment by consumer companies to match that investment by evaluating these solutions and conducting proof-of-concept projects with them.
- Continued focus, by both solution providers and consumers, on data security, compliance (including privacy), and data controls.

FURTHER READING

For more information on alternative analytical data flows, refer to the following resources:

- Gartner Logical Data Warehouse[12]
- The Data Warehousing Institute[13]
- Tech Target[14]

For more information on traditional data warehouse processing, refer to the following sources:

- Definition of data warehouse[15]
- Information Management[16]

[12] http://blogs.gartner.com/merv-adrian/2011/11/03/mark-beyer-father-of-the-logical-data-warehouse-guest-post
[13] http://tdwi.org/articles/2012/10/24/reconsidering-data-warehouse.aspx
[14] http://searchdatamanagement.techtarget.com/essentialguide/Big-data-applications-Real-world-strategies-for-managing-big-data
[15] http://en.wikipedia.org/wiki/Data_warehouse
[16] www.information-management.com/issues/19991201/1667-1.html