Data Abstraction Best Practices with Cisco Data Virtualization

Executive Summary

Enterprises are seeking ways to improve their overall profitability, cut costs, and reduce risk by providing better access to information assets. Significant volumes of complex, diverse data spread across various technology and application silos make it difficult for organizations to meet these objectives. To further complicate matters, there are a range of problems such as separate access mechanisms, syntax, and security for each source; lack of proper structure for business user or application consumption and reuse; incomplete or duplicate data; and a mixture of latency issues.

Data abstraction overcomes these challenges by transforming data from its native structure and syntax into views and data services that are much easier for application developers to use. Enterprises can approach data abstraction three ways: manual data abstraction, creation of data warehouse schemas, and data virtualization. Of the three approaches, data virtualization is the superior solution for data abstraction because it provides the most flexibility and agility to quickly retrieve data from different data locations and sources in real time.

Cisco® Data Virtualization is composed of different layers that form a data reference architecture that supports multiple consuming applications. The architecture aligns closely with analyst best practices mapped out by both Forrester and Gartner on the topic of data virtualization. This document explains data abstraction best practices using Cisco Data Virtualization that will enable your company to access the right data on demand, gain agility and efficiency, maintain end-to-end control, and increase security of your data across all your data resources.
Business and IT Challenges with Data Management

With large amounts of complex and diverse data spread across different application silos, enterprises are finding it difficult to gain access to their data. This large spread of diverse data also makes it difficult for enterprises to cut costs and reduce risk. A range of problems such as: separate access mechanisms, syntax, and security for each source; a lack of proper structure for business user or application consumption and reuse; incomplete data or duplicate data; and a mixture of latency issues demands a data management solution that can simplify data access. (See Figure 1.)

Figure 1. Data Abstraction Challenges

How Data Abstraction Overcomes These Challenges

Data abstraction overcomes source-to-consumer incompatibility by transforming data from its native structure and syntax into reusable views and data services that are easy for application developers to understand and consume.

Some data abstraction approaches enterprises use today include:

- **Manual data abstraction**: Some organizations manually build data abstraction in Java or use business process management (BPM) tools. Unfortunately, these are often rigid and inefficient. Such approaches are not effective for large data sets because they lack the robust federation and query optimization functions required to meet data consumers’ rigorous performance demands.

- **Data warehouse schemas**: Data modeling strategies for dimensions, hierarchies, facts, and other data organization methods are well documented. However, the data warehousing approach brings high costs and lack of agility. Also, data warehouse-based schemas do not include the many new groups of data (big data, cloud data, external data services, and more) that reside outside the data warehouse.
Data Virtualization Is a Superior Solution for Data Abstraction

Data virtualization is an optimal way to implement data abstraction for enterprises. From an enterprise architecture point of view, the Cisco Data Virtualization solution forms semantic abstraction, or a data services layer, in support of multiple consuming applications. The middle layer of reusable services decouples the underlying source data and consuming solution layers. This provides the flexibility required to deal with each layer in the most effective manner, as well as the agility to work quickly across layers as applications, schemas, or underlying data sources change. (See Figure 2.)

Data abstraction with data virtualization helps enterprises achieve a number of primary objectives, including:

- **The right business information at the right time**: Fulfill complete information needs on demand by linking multiple diverse data sources together for delivery in real time.

- **Business and IT model alignment**: Gain agility, efficiency, and reuse across applications with an enterprise information model or logical business model. Known as the canonical model, this abstracted approach overcomes data complexity, structure, and location issues.

- **Business and IT change insulation**: Insulate consuming applications from changes in the source and vice versa. Developers create their applications based on a more stable view of the data, allowing ongoing changes and relocation of physical data sources without affecting consumers.

- **End-to-end control**: Use a single platform to design, develop, manage, and monitor data access and delivery processes across multiple sources and consumers.

- **More secure data**: Consistently apply data security rules across all data sources and consumers with unified security methods and controls.

Cisco Data Virtualization Data Abstraction Reference Architecture

Figure 3 outlines the layers that form the data abstraction reference architecture. Architects and analysts can use this as a guide when abstracting data using the data virtualization platform.
The various layers included in this reference architecture are:

- **Data consumers**: Client applications need to retrieve data in different formats and protocols that they understand. Cisco Data Virtualization delivers the data to consumers using the most popular standards, including SOAP, REST, JDBC, and so on.

- **Application layer**: The application layer serves to map the business layer into the application format that each consumer wants to see. Examples include formatting into XML for web services or creating views with different alias names that match the way the consumers are used to seeing their data.

- **Business layer**: The business layer is built on the idea that the business has a standard or canonical way to describe primary business groups such as customers and products. In the financial industry, for example, information is often accessed according to financial instruments or issuers. Typically, a data modeler would work with business experts and data providers to define a set of “logical” or “canonical” views that represent these business groups. These views are reusable components that can and should be used across business lines by multiple consumers.

- **Physical layer**: The physical layer provides access to underlying data sources and performs a physical to logical mapping by integrating physical metadata and formatting views:
  
  - **Physical metadata**: Data that is essentially imported from the physical data sources and used to onboard the metadata required by the data abstraction layer to perform its mapping functions. As an “as-is” layer, group names and attributes are never changed in this layer.
  
  - **Formatting views**: These provide a way to map the physical metadata into the data virtualization layer by aliasing the physical names to logical names. The formatting views can facilitate simple tasks such as value formatting, data type casting, derived
columns, and light data quality mapping. This layer is derived from the physical sources and performs a one-to-one mapping between the physical source attributes and their corresponding “logical/canonical” attribute name. Also, this layer serves as a buffer between the physical source and the logical business layer views. Therefore, caching may be introduced at this level when it makes sense. Rebinding to different physical views during deployment is another role these views take on. Naming conventions are very important and introduced in this layer.

- **Data sources**: The data sources are the physical information assets that exist within and outside an organization. These assets may be databases, packaged applications such as SAP, web services, Excel spreadsheets and so on.

**Enabling the Forrester Data Virtualization Vision**

Forrester Research provides the following guidance for data abstraction in its “Data Virtualization Reaches Critical Mass” report.¹

Forrester says the most successful implementation of data virtualization uses a layered architecture that combines physical and virtual data stores at the appropriate levels to fit different performance requirements for different areas within the company. By funneling mappings of different source data through canonical business models, this creates an hourglass-shaped architecture.

Besides using canonical models in the middle of the architecture, there are two other important characteristics of effective data abstraction to note. First, physical data sources tend to be located more in the staging layers close to the actual data, whereas virtual data occurs more as the data gets closer to the end users. Second, a final virtual mapping layer provides data to consumers in the proper format. (See Figure 4.)

**Figure 4.** Forrester Research Data Virtualization Reaches Critical Mass

There is a striking resemblance between the Forrester and Cisco Data Virtualization best practice architectures shown in Figure 5.

---

Enabling the Gartner Discipline of Data Integration

Since 2005, Gartner has been researching the concept of data services in relation to the broader business and IT evolution. More recently, Gartner discussed the “discipline of data integration” at the Business Intelligence Summit, as shown in Figure 6:

Gartner says data integration includes the practices, architectural techniques, and tools used to gain consistent access to data, regardless of data structure type or group, in order to meet the requirements of applications and business processes. Data integration capabilities are an important part of an information-focused infrastructure and will drive the alignment and delivery of data to support BI and performance management. New challenges with data are creating a global surge of investment in data integration. Business factors such as the desire to increase speed to market or gain agility with business processes are causing organizations to manage their data differently. To accomplish these initiatives, companies need better visibility of their data in order to truly understand their performance and operations.

The data virtualization data abstraction reference architecture can be used to implement the Gartner “discipline of data integration” as follows:

---

• **Practices:** Cisco Data Virtualization has shaped the best practices that customers use today to implement data virtualization in their organizations, as well as influenced the practices recommended by leading IT analysts and system integrators. This thought leadership and real-world experience help users gain confidence when deploying data virtualization in their organization.

• **Architectural techniques:** The Cisco Data Virtualization Plan and Build Service brings a wealth of knowledge and skills to help users architect their data virtualization solutions. Architectural techniques are included in the service, which is designed to help customers get a project up and running quickly and maximize their return. Customers are introduced to the “Data Abstraction Best Practices Technical Guide,” which is used as an architectural techniques blueprint.

• **Tools:** The Data Virtualization platform provides a complete and proven tool to implement the Gartner “Discipline of data integration.”

• **Business context services:** In Data Virtualization’s reference architecture, the application layer provides the mechanisms for mapping and publishing views or web services in the context of the applications. The application layer maps into Gartner business context services. Application consumers require delivery of data using different protocols. Within the Data Virtualization reference model, data consumers use a variety of standard protocols, including JDBC, ODBC, SOAP/HTTP, REST and ADO/.Net to access needed data. These standard protocols support the BI, MDM, web service APIs, and enterprise objects consumers included by Gartner.

• **Semantic/logical services:** Gartner “semantic/logical” services provide for the transformation of the physical model into the business context view of the information. The terms logical and semantic are often referred to as canonical. It is a way of defining a common data dictionary across the business. The terms or attributes from this data dictionary are grouped together into semantically similar entities. Data Virtualization supports these needs with its formatting views.

• **Data manipulation services:** Gartner “manipulation” functions include access, storage, and delivery, which align with Data Virtualization’s physical layer. This is where introspection, discovery, and source data access tools expose the physical layer. Increasingly, Data Virtualization is providing access to a wide array of data sources, including relational, service oriented, file, packaged applications, and big data.

• **Optimization:** Both Gartner and Cisco view optimization as spanning the entire architecture from source to consumer, during both design and runtime, perfectly matching how Data Virtualization’s optimizers work.

Recent Gartner research on the logical data warehouse extends and enhances this guidance.

**Summary of Primary Benefits**

Data abstraction bridges the gap between the original form of business needs and source data. This best practice implementation of Cisco Data Virtualization provides the following benefits:
• **Simplifies information access**: Bridge business and IT terminology and technology so both can succeed.

• **Generates common business view of the data**: Gain agility, efficiency, and reusability across applications using an enterprise information model or “canonical” model.

• **Provides more accurate data**: Consistently apply data quality and validation rules across all data sources.

• **Provides more secure data**: Consistently apply data security rules across all data sources and consumers using a unified security framework.

• **Gains end-to-end control**: Use Data Virtualization to consistently manage data access and delivery across multiple sources and consumers.

• **Insulates business and IT change**: Insulate consuming applications from changes in the source and vice versa. Business users and applications developers work with a more stable view of the data. IT can make ongoing changes and relocation of physical data sources without affecting information users.

**Practical Next Steps**

Enterprises can begin achieving the primary agility and total cost of ownership benefits described earlier with a few simple steps. It is important to get started quickly with a manageable project that enables learning and a foundation for progress:

• **Set achievable goals**: Start with projects and a focused team. With success, broaden business and IT team involvement to expand usage across departments for ultimate full enterprise-level deployment.

• **Determine levels of abstraction**: Are the four recommended layers right for your organization? Do you need greater depth within one or more layers? The Cisco Data Virtualization Plan and Build Service can help answer these questions and get you started on the right path.

• **Determine modeling and mapping approach**: Should you use top down, bottom up, or some of both?
  
  o **Top down**: You have a vision, and you want to find the data to fulfill it. This is often referred to as contract-first design. In this approach Data Virtualization allows you to start with your own WSDL and map Data Virtualization services to your contract.

  o **Bottom up**: You know what your data looks like, but need to determine how you make it usable by others. In this approach, Data Virtualization allows you to generate or publish resources such as SQL view and web services directly from the Data Virtualization introspected sources.

  o **Both**: Mix and match appropriately according to domains and needs.
• **Start now:** Do not overanalyze. Getting started now with small steps is the best way to learn, progress, and gain value.

**For More Information**

To learn more about Cisco Data Virtualization, speak with your Cisco representative or visit [cisco.com/go/datavirtualization](http://cisco.com/go/datavirtualization).