PERSPECTIVE

DATA VIRTUALIZATION – A POTENTIAL ANTIDOTE FOR BIG DATA GROWING PAINS

Abstract

Enterprises are already facing challenges around data consolidation, heterogeneity, quality, and value. Now they must now also contend with yet another dimension – the Big Data. Let us examine specific challenges and one of the mitigation strategies, i.e., Data Virtualization.

Infosys®
Navigate your next

Introduction

Enterprises are already facing challenges around data consolidation, heterogeneity, quality, and value. Now they must also contend with yet another dimension – the Big Data. Let us examine specific challenges and one of the mitigation strategies, i.e., Data Virtualization.

Challenges faced by Businesses

Consolidating, organizing, and realizing the value of corporate data assets has been a long standing challenge. Various strategies have been in play for taking on this challenge and organizations are actively leveraging Operational Data Stores, Data Warehouses (DW), and Data Marts to deal with some of the requirements. However, the time taken to deliver tangible value has been a nagging problem in addition to the formidable resource commitment required to implement such systems. Latency of information for Business Intelligence (BI) and analytic use cases is another challenge since typical DW/BI processes and technologies mostly rely on batch processes to consolidate and present the data.

Enter Big Data – and we now have an even more challenging and nagging problem. Enterprises must deal with Variety, Velocity, and Volume of data hitherto unseen or even unknown. Dealing with traditional enterprise data in conjunction with big data has become a competitive necessity than merely a competitive advantage.

Another business scenario where all these challenges bubble up is Mergers & Acquisitions. In such situations, apart from the impedance mismatch on technology platforms, the same issue occurs on data semantics and core data models belonging to merged entities as well.

Opportunity cost for non-adoption

Not doing anything about data that is available in web logs, social media sites, and machine sensors is not an option. To remain competitive, enterprises must jump on the big data bandwagon but at the same time must be ready for newer sets of challenges.

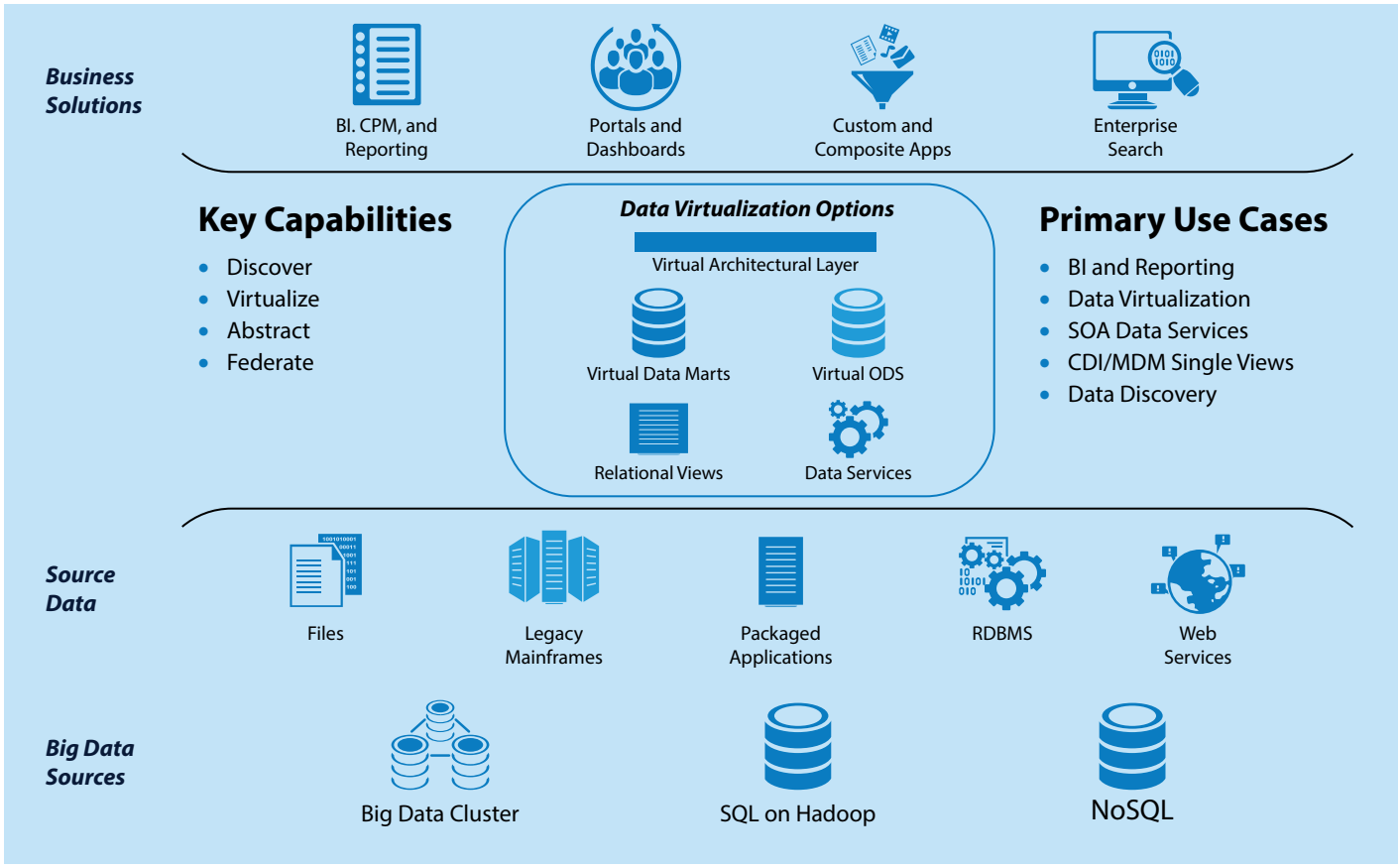
Being able to mash-up data from enterprise systems and big data platforms is an imperative. Many Use Cases revolve around the ability to combine data in a meaningful and timely fashion from all kinds of data repositories – conventional and new age. In the same vein, not being able to close out on rapid integration promises often dents the value propositions behind M&A deals.

Infosys view point on the area outlining how businesses can address the challenges

Heterogeneity of data environments is only growing and different classes of technologies must play together if organizations were to leverage a meaningful and consolidated view of enterprise data. Just consolidating across technology stacks doesn't quite solve the problem – an added dimension to worry about is the data location. Several data assets may be inside corporate firewall – but many other data sources reside in public and private clouds. Some may belong to the enterprise itself while others may be owned by partners, customers, and third-party data vendors.

A potential solution

Technology consultants and system integrators must constantly innovate in the area of data integration that impacts not only the operational efficiencies but also the reporting, BI, and analytical capabilities of an organization. The concept of data federation (that later evolved into Data Virtualization) has been around for a while. Lately, information fragmentation has rapidly increased on account of distributed technologies and new classes of data sources. Data Virtualization is now being looked at as one of the strategies to take on the challenges of information value *at the right time, at the right price.*



Let us consider a few real world scenarios where data virtualization can play a significant role:

- Real-time data integration to support fast-paced business processes and operational intelligence, that often require data from data warehouses to be enriched with data from operational systems
- Need for migration of data from a purely relational to a big data platform to enable a blend of structured and unstructured data to support business in the longer term
- Need to correlate vast amounts of data for customer 360° analytics where customer related data exists on a big data platform besides the ERP, and CRM type of systems that hold traditional transactions such as orders and warranty claims
- Need for a data-service-layer to serve up data from heterogeneous platforms including Hadoop based data repositories - to multiple downstream systems as well as to business partner systems
- M&A transactions requiring an immediate workable solution integrating multiple data repositories from two or more merged entities – without the usual lifecycle delays
- Data scientists needing to create analytics apps based on data in a big data lake, MDM systems, and ad-hoc data extracted from specialized systems

Informatica and Data Virtualization

There are a handful of vendors including Informatica that have made the area of data virtualization a priority, based on several well-known business imperatives. The Informatica Platform is a leading product in this space as per Forrester Wave Enterprise Data Virtualization 2015. With more than 5000 customers globally, a number of customers have indicated that they are leveraging Informatica for data warehouse augmentation, federation of master data, big data analytics, and real-time analytics.

Informatica's strong roots in data integration technologies position it as a top vendor that enables:

- Real-time data integration to support real-time DW and operational intelligence
- Data federation to provide basic multi-source data visibility
- Data virtualization to provide business friendly data abstraction, collaboration, increased data trust, and reusable data assets and services to be leveraged by composite applications
- Data profiling and quality services as part of data virtualization to increase the intrinsic value of data

Informatica Data Virtualization is a part of its wider integration platform and has the core capabilities of a Data Virtualization platform:

- Semantic Data Layer
- Role Based Access
- Data Caching
- Data Publishing (Batch, RESTful, Message Based, SOA)

Business Case

On the face of it, a virtualization approach looks like a no-brainer, however a business case must still be made to justify the introduction of yet another class of technology in the organization. There is no denying the fact that data virtualization is fast becoming a necessary pattern to realize the vision of an analytics organization that must sift through data - big or small; with or without time lags; with or without data quality processes; and with or without IT involvement.

Let us consider a hypothetical (but very realistic) scenario where a business division of a large enterprise:

- Must consolidate transactional data from multiple upstream processes both inside and outside company firewalls, with some of the sources existing in virtual private clouds – where data consolidation requires transformations that range from low to high complexity.
- Must blend and consolidate data on Hadoop (e.g. Cloudera), RDBMS (e.g. Oracle), and MPP (e.g., Teradata) platforms
- Must leverage master data from a number of Master Data Management (MDM) systems while also feeding these MDM systems with master data based on transactions
- Must provide multi-modal access to the data by providing SQL, NOSQL, Web Services, and MapReduce access.
- Must conform to SLAs surrounding data availability, performance (batch and real-time access), and scalability
- Must provide appropriate data access controls to manage different consumer populations under legal and company policy constraints

The following table captures a number of relevant dimensions to compare and contrast a Data Virtualization-based approach with more traditional approaches, where you bring in data into repositories meant for specialized usages such as reporting and analytics.

Dimension	Conventional Approach	Data Virtualization Approach	Comments
Infrastructure Resource Needs	High	Medium	Multiple data stores, more HW/SW infrastructure
Development Resource Needs	High	Medium	(data movement, quality checks, multi-level data modeling) vs. (virtualized consumption views)
Data Duplication & Quality Issues	High	Low	More hops, more chances of duplication and other quality issues
Data Governance Needs	High	Low	More governance required for data in specialized data repositories
System Coupling	Low	High	DV approach clearly results into tighter coupling that could slow down certain changes in the sources
Operational Impact on Data Sources	Low	Medium	Despite provision for data caching, etc., the impact on the source systems could be significant where the data velocity is high in those systems
Information Latency	High	Low	DV approach provides access to the data without delays inherent in ETL type of approaches
Consolidation Performance & Scalability Issues	High	None	High in the conventional approach vs. none in DV approach
Consumption Performance & Scalability Issues	Medium	High	Comparatively high in DV approach
Data Access Control Issues	High	Medium	Individual source systems can take responsibility of data access in the DV approach – while the conventional approach may have to replicate the access controls of the source systems
Production Support	High	Medium	More moving parts in conventional approach
Time to Value	High	Low	One of the main benefits of DV approach
Total Cost of Ownership	High	Medium	Total Cost of Ownership High Medium More pieces to build and manage in conventional approach

For the scenario described above, a virtualization approach seems to yield higher benefit, but the ultimate benefit realized can also depend on a number of other soft factors such as cohesion between data provisioning and data consuming teams, strengthening of data governance structures, etc.

In summary, a DV approach has become not just a viable strategy, but can also be construed as a necessity, given the market dynamics and requirements around innovation related to big data and simply time to value. The maturity and execution of bringing all the enterprise data – transactional, master, and reference – in a big data lake, is still growing but is nowhere close to where it needs to be. In the meantime, Data Virtualization technology can provide an effective bridge to the ideal future state.

About the Author

Atul Shrivastava

Principal Architect at Infosys.

Atul Shrivastava, Ph.D., is a Principal Architect at Infosys with 20+ years of experience in the IT Strategy, Architecture, Delivery, and Management. He has advised several fortune-100/500 clients in the areas of Information Management, Data Architecture, BI/Analytics, Enterprise Architecture, and Enterprise Integration. His interests include information value chain optimization, analytics driven business decision frameworks, and DW/ODS/SOA based information integration. He can be reached at atul_shrivastava@infosys.com.

For more information, contact askus@infosys.com

Infosys®
Navigate your next