Data Virtualization: Enabling Next-Generation Business Intelligence
Vijay Raj Singh

Vijay Raj Singh is the Center of Excellence (CoE) lead for the Business Intelligence and Performance Management (BIPM) practice within the Technology Excellence Group (TEG) at Tata Consultancy Services (TCS). Over an experience of more than 11 years, he has successfully anchored several customer engagements; helping organizations across industries design and develop solutions involving business intelligence tools and technologies. In his current role, Raj provides architectural guidance and conducts technical reviews for TCS' HiTech projects, assists the presales teams in proposal management, and supports the learning and development teams in organizational competency building. He holds a Bachelor’s degree in electronics and instrumentation engineering from the Visvesvaraya Technological University, Karnataka.

Shobhna Bansal

Shobhna Bansal is the Center of Excellence (CoE) lead for the Analytics, Big Data, and Information Management (ABIM) practice within the Technology Excellence Group (TEG) at Tata Consultancy Services (TCS). She has over 15 years of experience, through which she has been involved in numerous consulting and solution development engagements for some of TCS' leading customers, across industries. Shobhna also supports the various presales teams during the proposal management process. She holds a Bachelor’s degree in electrical engineering from the National Institute of Technology, Karnataka (NITK).
Abstract

Data is a vital asset for enterprises and is touted to become even more important in the times to come. Organizations rely on data warehousing (DW) to manage and mine gathered data and business intelligence (BI) techniques to draw insights for business support, but the capabilities of existing systems need to be augmented to accommodate the changing nature of data. Owing to globalization, digitization, and a rapidly changing economic climate, its volume and complexity are growing at a staggering rate.

There is a pressing need for organizations to effectively and efficiently manage data and transform it into valuable information in a timely manner. To profitably sustain in the dynamic business environment that exists today, HiTech companies are heavily investing in tools and techniques based on advanced analytics to derive actionable insights that enable timely and effective decision making. These new age systems should have the ability to tap into the existing and new data sources – structured and unstructured, and should be user-friendly with easy self-service options, while ensuring real-time access to the required information. To cater to these demands, the BI and DW systems need to evolve quickly and adopt modern data integration practices.

In this paper, we have discussed how data virtualization can enhance the traditional BI and DW systems, and provide enterprises with the required information – in a timely fashion – to meet the evolving business requirements with greater agility, and at affordable costs.
<table>
<thead>
<tr>
<th></th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data Virtualization: The Key to Agile Business Intelligence</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>Using Data Virtualization for Developing Next Gen Business Intelligence Systems</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>How Data Virtualization Based Systems Will Help Enterprises</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>Current State of Adoption and Future Outlook</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>Conclusion</td>
<td>10</td>
</tr>
</tbody>
</table>
Data Virtualization: The Key to Agile Business Intelligence

Business intelligence (BI) helps organizations make informed decisions pertaining to products and processes, as well as optimize the supply chain, thereby helping them to nurture successful relationships with customers and partners. It also drives greater agility and responsiveness with regard to global events and risks. Data warehouse (DW) helps in managing data integrity and transaction history, thereby ensuring information-centric needs of business stakeholders and regulators are successfully met. However, the very architecture of a data warehouse introduces redundancy into the system since the same data points are stored in different formats across various layers such as staging server, operational data store (ODS), base layer, and data mart.

As new sources of data emerge, businesses require flexibility and agility to integrate them with enterprise data. For HiTech companies, equipment like printing machines and storage devices are very useful sources, data from which, if integrated with enterprise data, can provide business critical information. It provides near real-time insights on device performance and its availability. Similarly, for the human resource function in a professional services industry, sourcing data from the social ecosystem can be useful in assessing a potential candidate, thereby improving the cost effectiveness of the recruitment process.

However, conventional methods of data integration follow a pre-defined process that usually takes a couple of days to sometimes weeks, to integrate newer sources of data. Additionally, the existence of multiple analytical systems such as an analytics sandbox, Hadoop or Big Data store, and so on, adds to the complexity. Enterprises therefore need to bring about a paradigm shift in the way they store, transform, and retrieve data.

To facilitate real-time decision-making processes encompassing the entire business ecosystem, organizations must take their BI capabilities to the next level. Basic building blocks of a next generation BI system are:

- **Rapid response analytics** engine to enable agile decision-making processes by means of an analytics sandbox, on-demand data integration, and self-service analytics tools
- **Data virtualization and federation** to combine disparate data sources – in real-time – to a single virtual data layer that provides integrated data services to applications
- **Big Data management** to integrate data—which may be structured or unstructured, internal or external, in batches or in real-time or by streaming, on premise or through the cloud—from different sources
- **Data visualization and discovery** to visually query the data, spot important trends and conditions, and set alerts and threshold values on dynamic data

Data virtualization is an important prerequisite to deploying the next-generation BI system, and offers a viable option to augment traditional data integration architectures. Going by emerging industry trends, data virtualization is a must-have component in every enterprise's data integration portfolio, more so for HiTech companies.
Data virtualization combines disparate sources, in real-time, into a single 'virtual' data layer that provides integrated data services to consuming applications.

In the era of Internet of Things (IoT), Big Data, and cloud, organizations tend to accumulate huge volumes of unstructured data, 'warehousing' which, is not viable. Data virtualization can harness value from all types of data to provide an integrated view in a standard format to downstream systems.

Using Data Virtualization for Developing Next Gen Business Intelligence Systems

A comprehensive data virtualization platform should provide an organization with a virtualized access capability to enable easy access to data from different sources. It should also help to collate data from diverse sources and present a unified view that is easily accessible. In addition, the platform should facilitate the abstraction of technical intricacies of data such as location, storage structure, accessibility, and storage technology, and allow manipulation, transformation, quality, and standardization of source data. Additionally, the platform should ensure a seamless delivery of information to relevant stakeholders, by publishing result sets in a standard format of open database connectivity (ODBC) views and data services, or web services such as simple object access protocol (SOAP), representational state transfer (REST), and XML, for on-demand consumption by a variety of data-consuming client applications.

To achieve the aforementioned capabilities, a data virtualization platform should essentially have the following components (refer Figure 1):

- **Data sources:** These refer to the structured and unstructured physical information assets that exist within and outside an organization. These sources may be databases, sensors, internet, cloud applications, packaged applications such as SAP, web services, data warehouse or data mart, and spreadsheets.

- **Data integration layer:** Lying at the heart of a data virtualization platform, this layer offers comprehensive data manipulation services by integrating data – in real-time – from a variety of sources as per predefined business rules. It also provides the ability to access, transform, profile, standardize, and federate data using an advanced query engine.

- **Business semantic layer:** This layer defines a set of logical or canonical views that represent a business entity grouped into subject areas, defined by the organization's enterprise information model. These views or models are reusable components that can be used by multiple consumers across the various business units in an organization.

- **Application layer:** This layer provisions the business semantics into formats customized to requirements of different end users. It can include formatting into XML for web services, or creating views under various aliases that match the way the users view data relevant to them. It also provides a mechanism to publish views, data, or web services, and to host these services for downstream consumption.
**Data consumers:** Application users require the delivery of data through different protocols that they understand, such as JDBC, ODBC, SOAP or Hypertext Transfer Protocol (HTTP), and REST. These standard protocols extend the consumption of data to a variety of downstream applications such as BI, master data management (MDM), spreadsheets, web service application programming interfaces (APIs), and other enterprise object consuming applications.

**How Data Virtualization Based Systems Will Help Enterprises**

Data virtualization greatly enhances the conventional BI/DW architecture improving implementation flexibility and business agility. Following are some important use cases that leverage data virtualization:

- **Data warehouse augmentation:** Data virtualization can significantly augment the capabilities of a data warehouse. With the increased uptake of digital technologies, a huge volume and a wide variety of data is generated in an organization’s ecosystem. For example, the transaction data generated from machines like printers provide key insights of customer usage patterns that can help fine tune the sales strategy, identify cross-
sell or up-sell opportunities, and strengthen key features that are heavily used by consumers. However, physical consolidation of such enormous volumes of ungoverned and unstructured data into a warehouse or a mart is not a feasible option. Therefore, data from multiple systems, tied with data services and metadata, will extend an existing data warehouse to become the ‘logical enterprise data warehouse’.

**Real-time data integration:** Real-time access to data from disparate sources provides faster and comprehensive business insights for building a competitive advantage. Hence, organizations need to extract data at the point of generation without waiting for it to get replicated to a central repository. To ensure real-time and current data for actionable analytics, a virtual layer can be created over the recently generated data in transactional systems, along with the historical data stored in data marts. HiTech companies like chip manufacturers and storage providers are exploring means to leverage real-time data generated by equipment for preventive maintenance and timely fault detection, thus enabling prognosis, diagnosis, and health management of machines.

**Data warehouse consolidation:** Data management programs conducted over a period of time, by decentralized organizations with relatively autonomous business units, almost always result in information silos. Enterprises realize that effective sharing of data across business units is vital to the smooth functioning of an organization. Data virtualization can help create an enterprise view by integrating siloed data warehouses and marts. In case of mergers and acquisitions (M&A), a data virtualization solution can provide an integrated view of customer as well as employee data of the participating companies, without having to wait for data warehouse consolidation.

**Self-service BI methods:** In the pursuit of agility and self-reliance, business users are no longer satisfied with just analyzing the given information. Users want to perform data discovery, exploration, and analytics with minimal support from IT teams. The data virtualization layer can facilitate this ‘self-service’ BI by creating data services with all the required information, which are then integrated with different data sources. Business users can spend more time in building their own reports and analytics rather than searching for the right information and trying to integrate it. This provides more flexibility, faster results, and involves less effort, leading to cost-effectiveness. Self-service BI is a highlight of the next-generation BI systems as organizations try to empower business users with governed data discovery capabilities as an expansion of the existing BI capabilities.

**Reuse across different BI platforms:** Enterprises often have multiple reporting tools in their BI landscape resulting in redundant and overlapping reports. The semantic or model created by different BI tools is limited to the tool’s own platform. Using data virtualization, a logical layer of canonical business entity model can be provided to enable consistent access by diverse downstream consumers across different business groups. It can render the semantics or models created by various data discovery and BI tools redundant. Business users can use any BI tool as desired, and still access the same data using consistent semantics. It also reduces the complexity and cost while migrating from one BI platform to another.

**Data acquisition platform:** In a typical data warehouse architecture, as data flows from the source to the mart layer, it creates various copies of redundant datasets. With data virtualization, we can minimize this redundancy by creating a virtual acquisition layer based on extracts of different source systems. A data virtualization platform has the ability to connect to a variety of sources with an API-based query interface, such as web
services (SOAP and REST), websites, and packaged applications like Salesforce. This layer acts as a single staging source for extraction, transformation, and loading (ETL) tools to extract data and process for downstream loads. It also eliminates the complexity of data acquisition process by abstraction of different data sources.

- **Prototyping and sandboxing:** Building a new data warehouse or enhancing an existing one are time-consuming exercises since the implementation team needs to follow a set of pre-defined tasks. Moreover, there is no process in place for the business team to validate the frequently changing requirements until the transformation is complete. This results in multiple iterations – something that can be successfully avoided by using data virtualization techniques. Data virtualization techniques help in prototyping requirements faster, thus taking relatively lesser time for solution development, and perform experimental analytics with different data sets, to realize timely benefits. The prototype thus developed can be validated in time by the business stakeholders, with the feedback getting incorporated in the subsequent iteration, thereby enabling progressive improvements. Moreover, one-time business requests that do not warrant the investment of either time or effort for a permanent solution, can be quickly accomplished through this methodology.

**Current State of Adoption and Future Outlook**

With growing information management challenges, companies are considering newer ways to integrate and manage data. Most organizations are therefore exploring data virtualization techniques to reap the numerous benefits of this approach that have been discussed in the previous section. Despite the increased focus on the data virtualization concept, the technical improvements, and growing use cases, the technology is in is still in earlier stages of adoption.

The growing information management challenges are compelling companies to consider newer ways to integrate and manage data – data virtualization is one such.

The adoption of the data virtualization technology depends on business requirements, with a special focus on future business needs. Most companies already have data management infrastructures in place and as a result, may be hesitant to make investments in this area. The use cases mentioned in the paper demonstrate the potential benefits of data virtualization based business intelligence systems. Hence, irrespective of the current state, businesses will soon make the shift towards an enhanced system.

To embark upon this transformation journey, organizations should keep the following points in consideration:

1. **Conceptualize:** Start with a manageable project that enables learning and lays a strong foundation for progress. Consolidate small-scale data sources into a single point of information to realize quick benefits. Define the guidelines, standards, and reference architecture for solution implementation.

2. **Develop:** Expand the usage across departments by broadening the involvement of IT and business teams. The existing BI/DW landscape can be optimized by:
   a. Implementing data virtualization to eliminate or reengineer redundant staging or acquisition data layers
b. Augmenting data warehouses and marts by integrating external data sources (like web, Big Data, social media, and so on) to improve the system's analytical capabilities

c. Replacing or reengineering data marts or entities as data services

3. **Operate**: Adopt data virtualization across the enterprise as 'Information as a Service' (IaaS) in a governed and an optimized way. Build logical canonical entity models for reuse across business units. Measure, monitor, report, and consume key metrics and information on a real-time basis.

**Conclusion**

With most organizations already having data management infrastructures in place, the adoption of newer technology faces an obvious roadblock – hesitance in making the humungous investment of time and effort. Data warehousing, over a period of time, has also proved its utility as the basis for an enterprise's decision support system, especially, in the following scenarios:

- When historical time series analysis of data across various dimensions is required
- When data quality is an issue, multiple passes and human intervention are required to fix it before loading it to a physical storage of a data warehouse
- When source systems are often offline due to maintenance activities or unreliable network connectivity
- When data integration requires transformation of huge volumes of data, which may cause performance issues if done in real-time

However, irrespective of current setups, the evolving needs of businesses will make it mandatory for technology to keep up with such transformations. With the emergence of in-memory computing, we are looking at the possibility of a single physical database that can cater to operational requirements, as well as analytical needs of a business. Data virtualization will play an important role in providing businesses with semantic and virtual models over the in-memory databases, as a platform for business intelligence based reporting.

Enterprise data warehouses will still exist, and data virtualization will augment – not replace – the traditional data integration systems. The scenarios mentioned in this paper demonstrate the potential of data virtualization to bring significant benefits to an organization, making this technology worth a serious consideration. Given the varying information needs, the quest for agility, integration challenges with new and emerging data sources, and advancement in technologies, data virtualization will be a key component in the overall data integration portfolio of an enterprise.
About TCS' High Tech Business Unit

Accelerated industry growth, rapid technological obsolescence, and the need for faster time to market compel High Tech organizations to improve business agility. High Tech solutions from TCS address fundamental industry problems, improve process efficiency, and enhance productivity and collaboration across businesses, while optimizing overheads.

From software product engineering and supply chain, to leveraging Internet of Things, digital re-imagination, cloud, Big Data, mobility and others, our solutions empower industry players - computer platform and services companies, software firms, electronics and semiconductor companies, and professional services firms to compete effectively.

Complementing our comprehensive service portfolio of IT solutions, business consulting, product engineering services, infrastructure services, and business process services, are our partnerships and alliances with leading industry vendors. Dedicated innovation labs, infrastructure support, and the Tata Research Development and Design Center (TRDDC) offer our clients access to cutting edge technologies, advanced systems engineering methodologies, storage optimization, and convergence solutions.

Contact
For more information about TCS' High Tech Business Unit, visit:
http://www.tcs.com/industries/high_tech/Pages/default.aspx

Email: HiTech.marketing@tcs.com

Subscribe to TCS White Papers
Feedburner: http://feeds2.feedburner.com/tcswhitepapers

About Tata Consultancy Services (TCS)
Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match.
TCS offers a consulting-led, integrated portfolio of IT and IT-enabled infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India’s largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com

IT Services
Business Solutions
Consulting