Hybrid Enterprise Data Lakes Provide Foundation for Disruptive Business Intelligence

Abstract

Enterprises are looking beyond traditional data warehousing practices to fulfill their business intelligence (BI) requirements. As the need to make accurate and timely decisions increases, enterprises seek real-time access to structured and unstructured data from multiple streams and logs.

A growing number of enterprises are exploring cloud and Big Data platforms to address this need. Moreover, since structured data warehouses and relational database management systems (RDBMS) are not enough to process large amounts of varied and unstructured information, businesses are looking to leverage hybrid data lakes, which are a combination of enterprise data warehouses and data lakes. Data lakes combine data existing in silos to improve information use and sharing, while lowering the overheads through reduced server and licensing costs.

The Need for a Fresh Perspective on Business Intelligence

Real-time decision-making is vital to achieve customer-centricity, improve brand image, and gain a distinct competitive advantage. This makes business intelligence (BI) and analytics a priority for C-suite executives across organizations. In today's information economy, enterprises need a scalable environment that helps them efficiently manage growing volumes of data, as well as handle diverse types of information needs. The findings of the TCS Global Trend Study 2015, 'Internet of Things: The Complete Reimaginative Force'[1] emphasize that businesses must be able to gather, process, and analyze huge amounts of digital data to realize the true potential of the Internet of Things (IoT). The speed of information retrieval is also critical to effectively leverage Big Data, compelling enterprises to embrace a new perspective on BI and analytics.

Adopting Next Gen Business Intelligence and Analytics

Data lake is a concept that has gained increased traction in recent times. It breaks down data silos, helping business analysts, data scientists, and engineers gain useful insights from the customer interaction data. The hybrid data lake, which is a combination of data lakes and data warehouses, offers several advantages over traditional data warehouses. Enterprises using the data lake can capture data from multiple customer touch points for in-depth analytics on customer psychographics and demographics. In a data warehouse environment, enterprises derive only descriptive and diagnostic analytics from the structured operational data stored.

Further, the development time needed to introduce a new data set to an enterprise data lake is much lesser than that needed for a warehouse . This ensures that the required data is available for analysis and reporting within acceptable timeframes. Data warehouses are also more expensive to maintain than data lakes owing to the high cost of hardware and licenses.

Additionally, the hybrid nature of the data lake allows easy and cost-effective analytics since summarized views from various data lakes can be seamlessly uploaded to the cloud platform for anytime-anywhere consumption by business users.

In essence, a hybrid enterprise data lake enables faster time-to-market through real-time data integration, schema-less data storage, and self-service BI and analytics. As a scalable solution, it drives real-time analytics, enabling enterprises to accelerate product and service innovation, and improve customer experience.

Optimizing Data Value with Hybrid Enterprise Data Lake

Traditional RDBMS and BI tools are not designed to handle the volume, velocity, and variety of Big Data. Organizations require a technology that integrates all types of data from internal and external sources to facilitate timely analysis. This will not only help them become agile, but also open up new business opportunities. Enterprises must therefore explore the advantages offered by cloud platforms and Big Data ecosystems. Various next-generation technologies allow distributed loading or parallel processing and analysis of large data volumes, enabling enterprises to scale quickly.

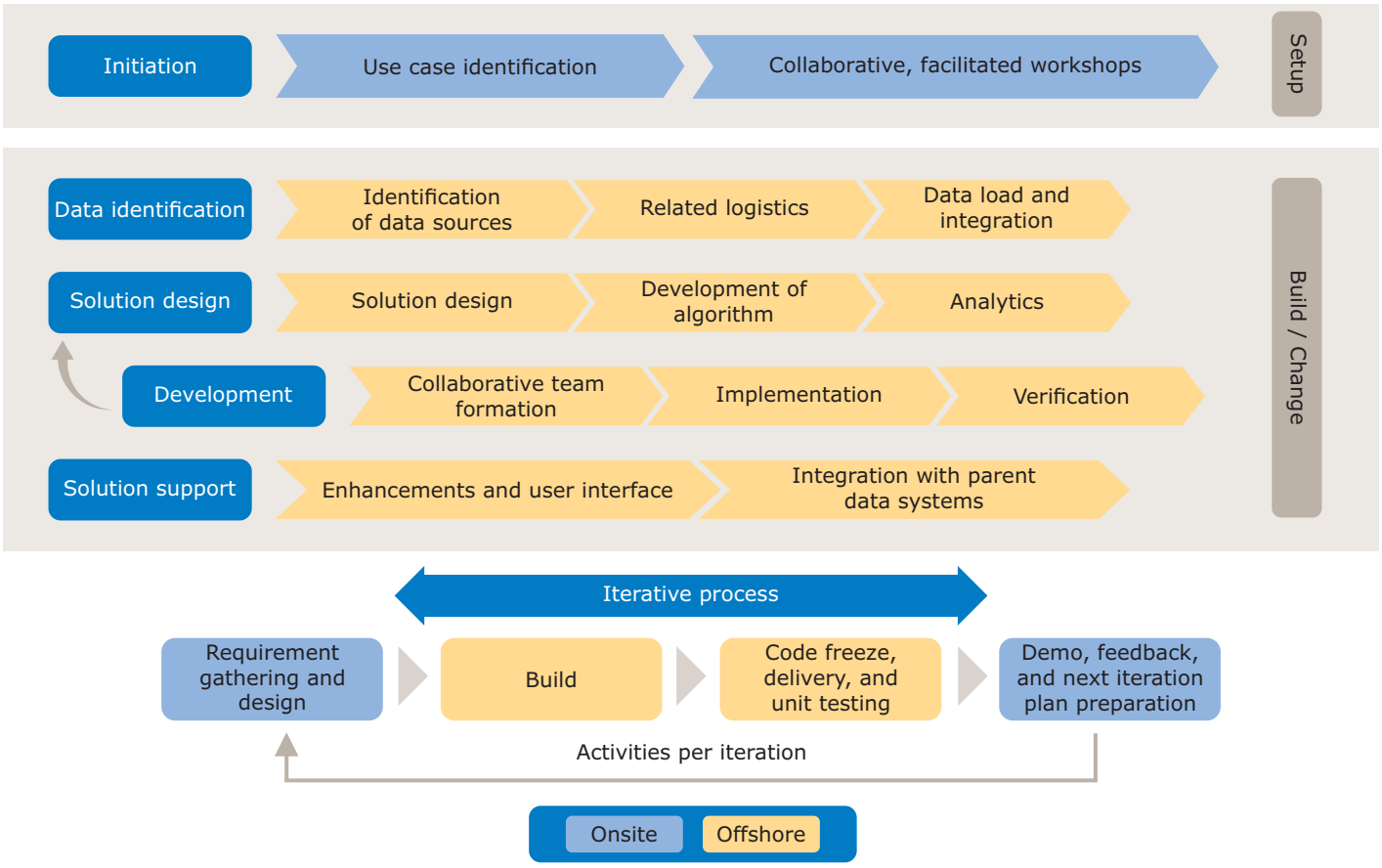
NoSQL databases are non-relational file systems that function in a distributed fashion over a cluster of commodity servers. They do not enforce a fixed schema, and are capable of supporting geographically-distributed architecture as well as large volumes of rapidly changing structured, semi-structured, and unstructured data.

A Hadoop platform, complemented by an MPP database, can be used for data lake implementation. Built on the 'shared nothing' concept, MPP allows efficient queries involving large tables. The entire operation comprises three steps: optimization, consolidation of results, and administration or coordination of activities among nodes. It enables real-time balancing of queries across nodes where large tables are split and stored. Each node processes the locally stored data, and coordinates with other nodes to consolidate and return the results. It runs every operation in parallel to provide the scalability required for Big Data systems. As a result, the fault tolerance and scalability achieved with MPP is very high.

While Hadoop can handle much of the workload on premise, pre-calculated and aggregated data can be hosted on the cloud in an MPP database. In some cases, a small-scale MPP engine can reside on premise for the calculation and aggregation of high volume analytical queries before it is uploaded on to the cloud for consumption by business users.

The constituents of such an ecosystem include the Hadoop Distributed File System (HDFS), NoSQL and columnar databases, Massive Parallel Processing (MPP) appliances, and in-memory processing.

The four categories of NoSQL databases widely used in data management systems are document databases, graph stores, key-value stores, and wide-column stores.



A roadmap to guide the implementation of an enterprise data lake

The company's vision, mission, goals, and key success factors should be taken into account in order to come up with the high-level business requirements. These requirements should cover data ingestion and access, security, data discovery and delivery, analytics, and data management, archival, back-up, and recovery. After a thorough analysis of the as-is system including existing data warehouse, sandbox, and data marts, and identification of areas of enhancements, the proofs-of-technology should be developed. The final steps should involve the assessment of information and program governance, covering information access and security; evaluation of costs; and the definition of data warehouse program organization model. Once the roadmap is ready, an agile and modular approach can help achieve quicker and more meaningful results with iterative development and shorter release cycles.

Business Benefits of Hybrid Enterprise Data Lake

Although a hybrid enterprise data lake involves complex data management architecture, it offers compelling business benefits:

- It enables enterprises to overcome the performance and functional limitations of existing legacy systems.
- Enterprises can enhance business decision support at various levels by integrating data from various data silos to respond faster to current and future regulatory and compliance requirements.
- The hybrid enterprise data lake generates strategic reports in response to standard and customized queries, in a cost-effective and timely manner.
- A hybrid enterprise data lake is capable of figuring out hidden patterns in customer data to help devise targeted marketing strategies.
- It helps enterprises gain better understanding of customers by capturing and integrating their behavioral and psychographic information from a variety of customer interactions.
- A hybrid enterprise data lake provides analytical capabilities not available within a traditional line-of-business application.
- It enables self-service for business groups, reducing the dependency on internal IT teams, and ensuring higher satisfaction among users.
- A hybrid data lake can also support other enhancement initiatives such as web personalization.

Conclusion

Hybrid enterprise data lakes help enterprises perform advanced analytics on large volumes of Big Data. Further, a hybrid environment enhances business reporting and analysis by mapping transactions and records to demographical, historical, and other contextual data, and paves the way for disruptive business intelligence.

While creating hybrid systems, enterprises must pay attention to data security, access control, regulatory compliance, and audit trail management. Without information governance, a data lake could end up being a collection of information silos in one place. Despite these challenges, when implemented correctly, hybrid enterprise data lake systems can create disruptive possibilities for maximizing information intelligence, enabling enterprises to realize greater returns on their BI investments.

References

[1] TCS, "Internet of Things: The Complete Reimaginative Force" (August 2015), accessed April 22, 2016, http://sites.tcs.com/internet-of-things/

About The Authors

Shobhna Bansal

Shobhna Bansal is a Solution Architect with the Digital Enterprise Transformation Group of the High Tech business unit at TCS. She has over 16 years of experience, and is the Data and Decision Science Practice Lead for the group.

Samrat Mukherjee

Samrat Mukherjee is an Enterprise Architect with TCS' Alliance and Technology Unit. He has over 14 years of experience in the area of enterprise information management and integration, and has worked on large-scale data integration and analytics projects comprising diverse technologies.

Contact

Visit the Communication, Media & Technology pages on tcs.com

Email: HiTech.Marketing@tcs.com

Subscribe to TCS White Papers

TCS.com RSS: http://www.tcs.com/rss_feeds/Pages/feed.aspx?f=w
Feedburner: http://feeds2.feedburner.com/tcswhitepapers

About Tata Consultancy Services Ltd (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

TCS Design Services I M I 11 I 16