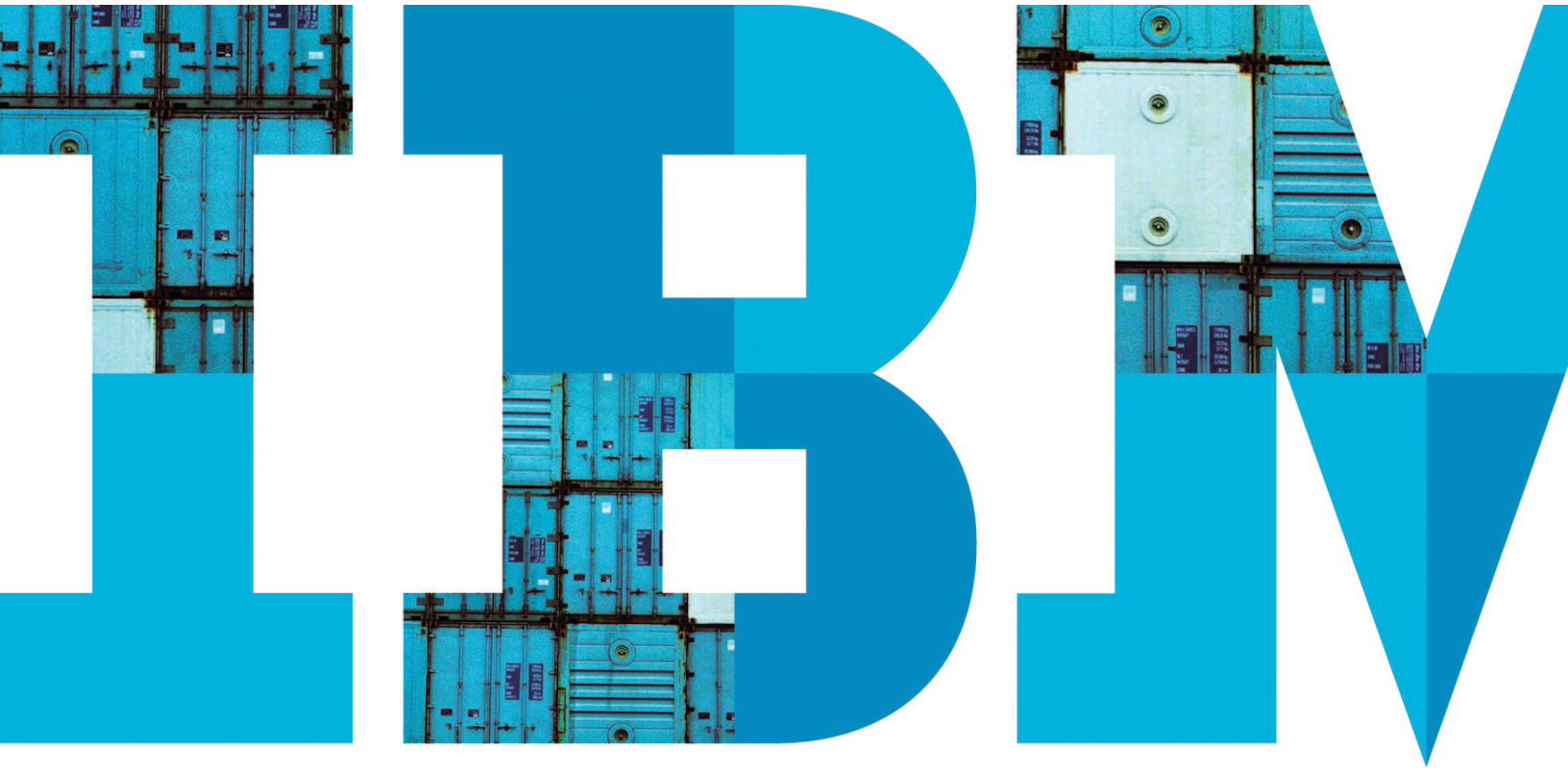
IBM Software
Thought Leadership White Paper

April 2014

Derive actionable real-time insight from your big data with data replication



Big data is big news. While organizations continue to explore and experiment with their information resources, they are now focusing on using big data technologies to solve real business problems. For those that have made this shift, IBM has identified patterns of usage for their first big data projects and has distilled them into several high-value use cases.

Five high-value uses for big data

IBM has conducted surveys, studied analysts' findings, talked with more than 300 customers and prospects, and implemented hundreds of big data solutions. Based on this research and experience, it has identified five high-value use cases that serve as excellent entry points for organizations embarking on a big data journey:

1. **Big data exploration:** Find, visualize and understand big data to improve decision making
2. **360-degree view of the customer:** Enhance the existing customer record by incorporating information from internal and external sources
3. **Security/intelligence extension:** Reduce risk, detect fraud and monitor cybersecurity in real time
4. **Operations analysis:** Analyze a variety of machine data for better business results and operational efficiency
5. **Data warehouse modernization:** Integrate big and traditional data warehouse capabilities to gain new business insights while optimizing the existing warehouse infrastructure

These entry points are independent, departmental and focused, so they do not have to be implemented in any particular order. It doesn't matter where an organization starts; it only matters that it starts.

The key is to identify which use cases make the most sense given the organization's current challenges. This paper focuses on data warehouse modernization, including the need to modernize, the technologies available and the potential advantages associated with this entry point.

"Data types and classes of data usage in data centers are proliferating like never before. As enterprises grapple with Big Data sources, including streaming data, and seek to blend and reconcile that data with analytical data in data warehouses and up-to-the-minute transactional data, the pressure is on to bring these elements together in a timely, yet organized manner for intelligent management and in support of real-time analytics and decisions at the point of action."

—Carl Olofson, IDC[1]

The drive toward data warehouse modernization

Industry experts have spotted a trend: organizations are moving away from a monolithic enterprise data warehouse (EDW) to a distributed architecture. But big data does not spell extinction for the data warehouse. On the contrary; big data problems can be better addressed by an enhanced, or modernized, data warehouse.

"[There's a belief] that if you want big data, you need to go out and buy Hadoop and then you're pretty much set. People shouldn't get ideas about turning off their relational systems and replacing them with Hadoop."

—Ken Rudin
Head of Analytics, Facebook[2]

In support of this middle-ground approach, there are two main drivers for data warehouse modernization. Organizations need to:

- Leverage a variety of data for business insights
- Optimize the warehouse infrastructure

The growing volume of constantly changing data challenges organizations to make informed, real-time business decisions and stay ahead of the competition. Savvy organizations are quickly realizing that they can extend the value of their existing systems and generate significant business advantages by capturing incremental changes in key data and delivering that information, in real time, to optimize data warehouses and business intelligence initiatives. This enables them to react more quickly to changes in customer sentiment, uncover new market opportunities and introduce groundbreaking new products aligned with the latest trends.

The next-generation architecture to support middle-ground, data warehouse modernization requirements (and real-time data processing and analytics) takes a leading position within this new architectural view (see Figure 1).

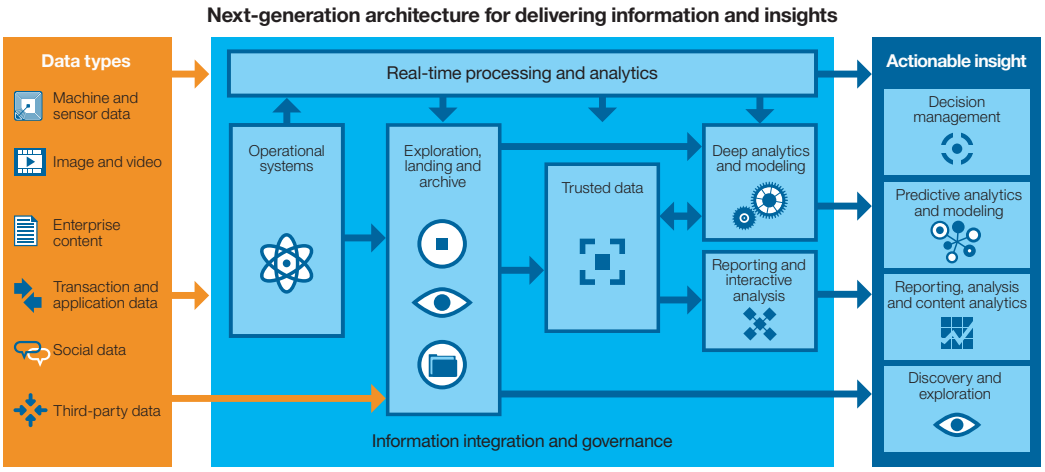


Figure 1. The IBM architecture for delivering information throughout an organization involves several zones that help transform data into actionable insight.

Real-time processing and the role of data replication

Data replication has evolved from the traditional data transformation and movement associated with batch and bulk data movement. Batch and bulk data movement is scheduled on a relatively infrequent basis for all data; real-time data transformation occurs only on changed data. The changed data is captured, transferred and transformed, and then applied to the target based on change data capture (CDC).

Data replication's real-time operational and analytical data synchronization enriches mobile applications and big data projects with up-to-the-second information. It can also be used to enable continuous availability across the data center or around the world. In heterogeneous environments, data replication supports data distribution and synchronization for transactional systems to support confident, up-to-the-minute decisions at the point of impact. Alternatively, when deployed in homogeneous environments, data replication supports business continuity and disaster recovery. In all scenarios, data replication helps minimize the cost of infrastructure and optimizes resource utilization.

Leveraging advanced data replication technology for accurate, real-time insights

IBM® InfoSphere™ Data Replication provides trusted data synchronization and availability, enabling organizations to efficiently and flexibly manage their big data growth and use current information to increase revenue. InfoSphere Data

Replication delivers high volumes of (big) data with very low latency, while providing the broadest and deepest support for sources, targets and platforms, ensuring the right information is available to augment data warehouses, data marts and point-of-impact solutions.

InfoSphere Data Replication is a critical component of IBM Watson™ Foundations, the IBM platform for big data and analytics. It can work in conjunction with technologies such as InfoSphere Streams (which performs in-motion analytics on a wide variety of relational and non-relational data). It also is tightly integrated with InfoSphere BigInsights™, bringing enterprise-class, real-time data synchronization and data availability capabilities to Apache Hadoop.

Many organizations are moving from a single data warehouse to a distributed architecture that includes both traditional and big data marts in a hub-and-spoke topology. Ensuring delivery of up-to-the-minute data across the entire environment becomes very important as organizations increasingly leverage Hadoop to cost-efficiently take on more of the traditional enterprise workload.

With the pre-built integration capabilities in InfoSphere Data Replication, Hadoop can be treated just as any other data target. InfoSphere Data Replication can deliver real-time data from multiple sources across the enterprise to mainframes and distributed platforms—and by using its Hadoop Distributed File System, deliver directly to Hadoop distributions such as InfoSphere BigInsights.

Improving visibility into lines of business

As data volumes grow, so does the time and processing it takes to perform traditional—and often redundant—extract, transform and load (ETL) processes against that growing data. To prevent increasingly long batch windows and subsequent service-level agreement (SLA) misses (or running into core business hours), businesses can use CDC techniques to support these ETL processes.

For businesses with large volumes of daily changes that can't afford downtime, CDC offers even more visibility into the data warehouse.

Instead of issuing SQL to extract data from database tables in a pull process, CDC capabilities can work against transaction logs in a push manner, thus lessening the impact on source systems while allowing transformation and load processes to be used only on new or changed data, in real time. If only simple (row-level SQL) transformations are required, in some cases those jobs may be replaced by end-to-end CDC jobs. The end-to-end CDC jobs may also replace large, nightly batches with intraday microbatches fed by CDC.

Case in point: Petroleum refiner reduces batch windows

For example, InfoSphere Data Replication helped a petroleum refiner optimize its custom batch processes. The batch window was beginning to significantly impact the running of its business. By using InfoSphere Data Replication, the refiner realized the following benefits:

- **Eliminated much of the legacy batch extraction code:** By leveraging "in-flight" transformation capabilities of InfoSphere Data Replication, IT removed much of the custom batch extraction code required to update the EDW. This also reduced development and maintenance costs.
- **Eliminated file locking issues during extraction:** InfoSphere Data Replication uses the native database transaction log to access transactions that have changed versus reading database tables directly. This eliminated file-locking issues that occurred during batch data extraction and allowed the petroleum refiner to extend its ERP application to business users.
- **Reduced batch window:** By significantly reducing nightly data extraction requirements, InfoSphere Data Replication helped the petroleum refiner reduce its batch window requirement from 15 hours to 6 hours while continuing to meet EDW SLAs with yearly data volume growth.

Combining InfoSphere Data Replication with InfoSphere Information Server

In the petroleum refiner's use case, the company was using custom ETL. Pairing InfoSphere Data Replication with an enterprise-ready ETL solution such as IBM InfoSphere Information Server, however, offers the best of both worlds, including tight integration and metadata lineage across products. InfoSphere Data Replication with CDC provides a noninvasive, reliable, low-impact approach for extracting changes from mission-critical systems and delivering the stream of incremental data changes to InfoSphere Information Server. This enables businesses to continuously update the data warehouse without requiring batch windows that involve transferring entire data sets. InfoSphere Data Replication can supply active data warehouses with continuously captured data, giving businesses fresh, up-to-date information for time-sensitive decision making and analytics efforts.

Increasing operational business intelligence while reducing CPU utilization

Operational business intelligence (BI) systems can increase visibility into lines of business, but implementation must be handled carefully. Directly querying mission-critical systems for reporting purposes places a heavy burden on those systems and results in increased CPU utilization, which may hamper application performance. Data replication helps increase availability of enterprise data for operational BI without negatively impacting source systems.

By replicating live production data to a secondary system (operational data store or enterprise data warehouse) for reporting and query requirements, InfoSphere Data Replication helps reduce costs and risk by avoiding impact to CPU utilization on mission-critical systems, which preserves application performance without affecting end users.

Case in point: Scotiabank modernizes its data warehouse

Scotiabank, a leading Canadian multinational financial services provider, undertook a data warehouse modernization initiative to reduce the time and cost of providing enhanced payment services.

The organization used InfoSphere Data Replication CDC capabilities to synchronize its online transaction processing (OLTP) and core banking databases with its data warehouse in near-real time. Their efforts have paid off handsomely:

- **Over 400,000 business banking monthly account statements are now delivered online**
- **The time required to retrieve data to create complex reports has decreased by 99 percent**
- **Report generation is tuned so that 95 percent of all reports are delivered on demand, negating the need to "pre-generate" reports**
- **Overall system performance is 100 to 200 percent faster across all report formats, increasing customer satisfaction**
- **The legacy system was re-engineered through CDC, creating transparency at the core system business logic layer that enables new employees to develop in a very short time (one to two months)**

To learn more about Scotiabank's data replication initiatives, visit: http://ibm.co/1crRovm

Beginning the big data journey with IBM Watson Foundations

The five big data use cases described in this paper, including data warehouse modernization, provide high-value starting points for companies looking to begin their big data journey. These organizations require an integrated set of technologies that are specifically designed to address the unique challenges of working with high-volume, high-variety and high-velocity data. These are not single-issue problems with single-product solutions.

IBM Watson Foundations, including InfoSphere Data Replication for real-time data delivery and augmentation, can play an integral role in that transformation. It provides a valuable foundation that helps you reduce the time and costs of big data projects, as well as achieve a rapid return on investment (ROI), by leveraging pre-integrated components. By building on those capabilities, you can start small with an initial use case and easily progress to others as you continue on your big data journey.

For more information

To learn more about IBM big data integration capabilities, including InfoSphere Data Replication, please contact your IBM representative or IBM Business Partner, or visit the following website: **ibm.com**/software/data/bigdata/use-cases.html

To learn more about big data use cases and the IBM big data platform, visit: **ibm.com**/software/data/bigdata/use-cases.html

Additionally, IBM Global Financing can help you acquire the software capabilities that your business needs in the most cost-effective and strategic way possible. We'll partner with credit-qualified clients to customize a financing solution to suit your business and development goals, enable effective cash management, and improve your total cost of ownership. Fund your critical IT investment and propel your business forward with IBM Global Financing. For more information, visit: **ibm.com**/financing

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
April 2014

[1] Olofson, Carl. "The Golden Age of Data Integration." IDC Link. IDC. December 23, 2013. Doc # lcUS24555813

[2] Lopez, Isaac. "Rudin: Big Data is More Than Hadoop." Datanami. Oct. 30, 2013. www.datanami.com/datanami/2013-10-30/rudin:_big_data_is_more_than_hadoop.html

Please Recycle

IMW14758-USEN-00