IBM Data Replication for Big Data

Low impact, log-based captures from source systems enable near real-time, incremental delivery of transactional data to Hadoop- and Kafka-based data lakes and data hubs.

Highlights

- Stream changes in real time giving your data users access to the most current and accurate data in your Hadoop- or Kafka-based data lakes or data hubs.
- Provide agility to data in your data warehouses and data lakes through a scalable solution with high volume throughput and low latency.
- Achieve minimum impact on source systems using log-based captures.
- Replicate data regardless of where it resides through support for a wide variety of sources and targets.

Big data technology platforms are providing business professionals with greater insights so they can make better decisions. Their success in extracting value from data is compelling leaders to unlock trusted data assets throughout their entire organization. To achieve this, data warehousing is being modernized, and data architecture is being improved. What used to be a single data warehouse node is now shifting to a distributed architecture.

Traditionally, it has been a challenge to make real-time data available to enterprise data hubs or data lakes. And to remain flexible and agile in the big data world, enterprises need to capture information with low impact to source systems. They also need to deliver changes to analytics and other systems at low latency, and analyze massive amounts of data in motion. The IBM® data replication portfolio helps capture and deliver critical dynamic data across an enterprise to expedite better decision making.

Making use of low impact, log-based data capture from transaction systems, IBM data replication delivers only the changed data across the enterprise. This enables organizations to capitalize on emerging opportunities and build a competitive advantage through more real-time analytics. Automation of this process unlocks a multitude of data stores to the business- and customer-facing platforms that are critical to success in the dynamic and connected environments that organizations are building. By providing a flexible, one-stop shop for trusted heterogeneous information replication, IBM data replication synchronizes transaction data with:

- A relational database based data warehouse, data mart, or operational data store.

- An Apache Hadoop data lake or an Apache Kafka landing zone or streaming hub
- A cloud data store
- Transformation engines such as IBM InfoSphere DataStage®.

IBM data replication also supports homogeneous scenarios, such as version-to-version database migration, maintenance of active/stand-by and active/active data environments for high availability, and database mirroring for workload balancing across systems.

Kafka and Hadoop target engines are powered by the standard replication architecture and capabilities: fault tolerance, scalability, security, and data lineage and auditing.

Additionally, access to IBM expertise and collaboration with big data domain leaders, such as Hortonworks, ensures that clients are provided with industry leading support.

IBM data replication

Moving data around the enterprise can be a challenging task, as enterprise environments comprise a variety of operating systems, databases and other data stores. The IBM data replication solution has been deployed by a variety of enterprise clients in the most challenging implementations. The solution offers proven scalability and performance while providing high-quality data to a wide variety of target systems with low impact to day-to-day activities taking place on the source or target systems.

Non-intrusive to applications and databases

Enterprises expect their chosen replication tool to be non-intrusive to applications and databases so that applications can continue normal operations while replication runs continuously. IBM data replication captures data from database logs (not from triggers or selects from tables). This ensures that the performance of even the most demanding mission-critical applications that are running on the source system are not adversely affected. *Figure 1*.

Ready for enterprise scale operational demands

Recognizing that enterprise clients expect near lights-out operations, IBM provides standard GUIs for centralized operation and ease of use. IBM data replication provides unparalleled scripting and API support for all configuration and monitoring needs, thereby ensuring that configuration, deployments, alert monitoring, and problem resolution can be automated.

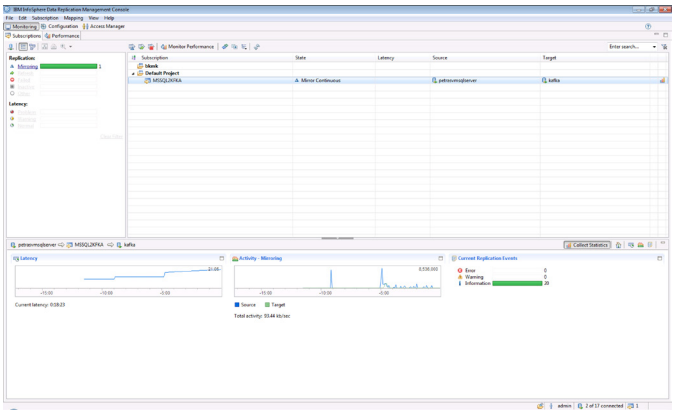

Figure 1: Shows the monitoring dashboard in the Management Console GUI.

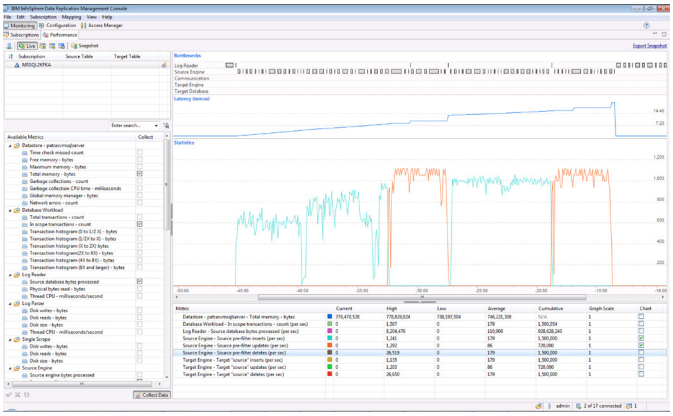

Figure 2: Monitor performance from the Management Console.

Industry standard authentication mechanism and security

Security for user access is provided via LDAP-based authentication, ensuring that industry standard security best practices are supported. Users can optionally choose to have an LDAP server manage their user credentials, user authentication, and data store access information. They can also rely on built-in capabilities available in the replication technology components to authenticate users, and store user credentials and data access information. Security of data at rest or in motion is assured via documented best practices for use of pervasive IT practices (for example, use of VPNs to secure data transmitted over a network).

Maintain transactional consistency

Transaction consistency is for the preservation of units of work and to maintain referential integrity. The software supports full transaction granularity with before-and-after images of all transactional changes. In addition, built in fault tolerance capabilities allow organizations to easily recover to the last committed transaction.

Data lineage support

Users can choose to preserve data lineage and available metadata associated with database changes. Using such capabilities, before-and-after images of record updates or metadata (such as the user who made the change in the original transaction) can be replicated and saved on the replication target. *Figure 2*.

Consistent performance and scalability

The IBM data replication solution is proven technology that has been successful deployed by a wide variety of enterprise clients. These users have repeatedly proved the technology's scalability and performance in the most challenging implementations. Additionally, incremental changes can be delivered for better business agility with very low latency to the target systems.

IBM data replication provides performance and scalability while supporting replication end points on a broad range of operating systems and all major databases, Kafka, Hadoop and more.

Real-time integration using Hadoop and Kafka

As organizations consume and drive insights with big data, it becomes necessary to merge operational transaction data (such as customer, product, and accounting data) with high-volume data streams (smart devices, social media, and web interactions). The IBM data replication portfolio provides this integration by targeting Hadoop and Kafka.

Feeding Apache Hadoop clusters directly via the data replication target apply for Hadoop

A Hadoop distributed file system (HDFS) under the control of a commercial Hadoop platform has become the de facto standard for the enterprise data lake. Built to process large, relatively static data sets with "bulk append" only, the HDFS file system is designed to distribute copies of data across commodity nodes to provide availability and scalability. Built from the ground up for high performing access in service of analytical applications and information exploration, Hadoop is ideally suited to the storage and analysis of vast amounts of unstructured data.

A typical enterprise will use the Hadoop platform to consolidate structured business information with high volume and unstructured data from social media, internet activity, sensor data and other sources. IBM data replication supports this strategy by delivering real-time feeds of transactional data from mainframes and distributed environments directly into Hadoop clusters. It accomplishes this via its change data capture (CDC) Hadoop target engine using a WebHDFS interface.

Analytics software programs that use the Hadoop clusters as their system of reference are then able to benefit from the unlocked transactional data assets made available by data replication. They result in better predictive, real-time and advanced analytics insights to power more agile and accurate business decision making or customer interaction.

However, to get dynamic system of record or engagement data into HDFS requires that the writer self-buffer on the way in to build large bulk load text-based flat files. Ongoing maintenance of the file system such as combining or removing files to maintain a healthy environment (for example, avoiding millions of files) is also required. Performance and availability of the HDFS cluster is directly impacted by the maintenance or lack thereof.

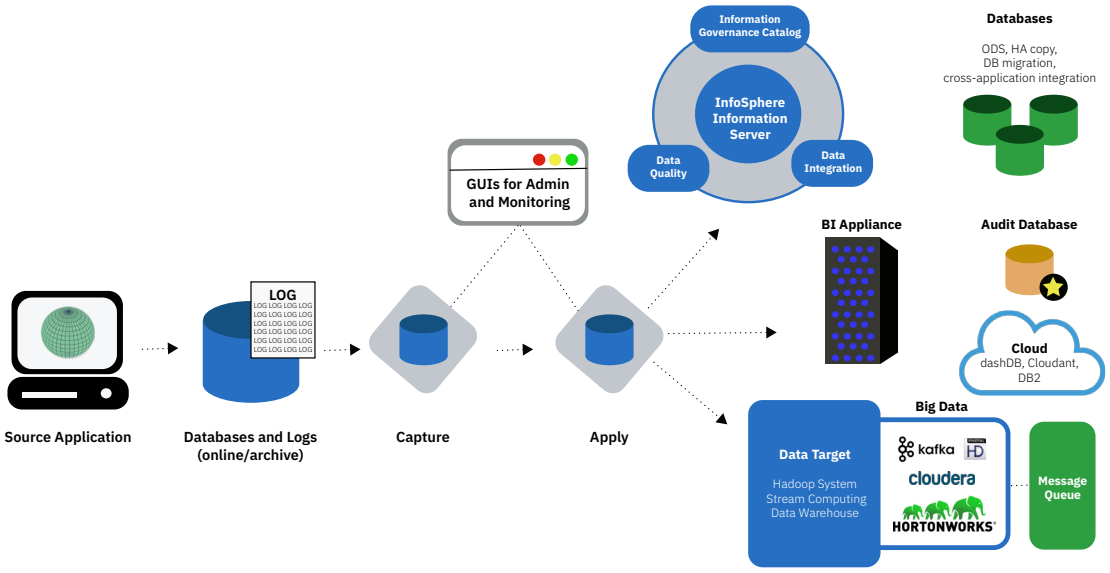

Figure 3: IBM data replication IIDR (CDC) Log Based Change Data Capture with real time feed to Kafka or Hadoop clusters and many other targets.

Additionally, files in HDFS have fixed formats such as a comma separated fields. There is no inherent schema data associated with Hadoop and HDFS. Readers of the data have to *agree* with writers on the field list and format. While there are best practices evolving, and HIVE metadata can be added to help with these limitations, it is often non-standard work in progress that varies highly from organization to organization and from technology to technology.

Expanding the data lake deployment with Apache Kafka

Apache Kafka provides a commodity hardware-based clustered data environment designed for real-time data streaming, processing and storage for structured, semi-structured and highly volatile data. Apache Kafka was designed from the outset to deal with constantly changing events and data. Its unique combination of low cost deployment, scalability with built in metadata handling capabilities, and self-managed storage compression is driving its growth as an information hub feeding the data lake and other big data environments.

Apache Kafka incorporates built in "insert with log compaction" to emulate deletes and updates. Storage is typically self-described JavaScript Object Notation (JSON) documents wrapped in Apache Avro binary formats. This introduces a metadata driven schema concept that is ideally suited to more structured data, such as that found in traditional transactional data sources, without compromising the ease of analysis and scalability and cost profile associated with the data lake. In effect, a Kafka *topic* conceptually represents a logical table or file definition of a traditional transactional data store.

Additionally, Kafka has built in facilities to maintain:

- Automated *compaction* to control and minimize administrative and storage costs.
- Non-keyed data using a sliding window of most recent data (for example, seven-day window, one-month window).
- Keyed data providing access to the record with the most recent value for each key.

Replicating to Kafka when Kafka is being used as a data hub or landing zone

In support of this strategy, IBM data replication provides a Kafka target engine that streams data into Kafka using either a Java API using a writer with built-in buffering, or a Representational State Transfer (REST) API that provides batch message posts. Thousands of topics (mapped to source tables) are easily handled and each can sustain millions of messages that represent the result of an insert, update, or delete on a transactional source.

With the data now amassed and available in self-described Kafka topics, Kafka consumers can feed data to the desired end points such as: IBM InfoSphere® Information Server, Hadoop clusters such as Hortonworks, and cloud-based data stores such as IBM Db2® Warehouse for Cloud or Db2 Hosted.

Improving the flexibility of message delivery into Kafka with Kafka custom operation processors

When messages are landed in a Kafka cluster, a variety of available connectors or consumers can retrieve these messages and deliver them to target destinations such as HDFS, Amazon S3 and Teradata.

Users can save time and costs by using available consumers, such as those listed on Confluent and Hortonworks.

The Change Data Capture (CDC) target replication engine for Kafka writes Kafka messages that contain the replicated data to Kafka topics. The replicated data in the Kafka messages is by default written in the Avro binary format. Consumers who want to read these messages from Kafka clusters needed to utilize an Avro binary deserializer.

To help clients solve this challenge, IIDR (CDC) now provides Kafka custom operation processors (KCOP) to improve the flexibility of message delivery into Kafka. Clients can make use of a number of integrated predefined output formats, or adapt these user exits to define their own custom formats, what data is included in the message payload, and more. In addition to giving users the flexibility of defining the message formats and payloads, users are now also able to specify the Kafka topic names for their message destinations.

Moreover, as more common clients needs are assessed, IBM will add more such predefined output formats.

Retrieving data with transactional semantics from Kafka using the IBM data replication CDC target engine

With IBM CDC, clients can leverage the performance and low cost of Kafka to have database-like transactional semantics without compromising performance when delivering changes to Kafka. Duplicate data is avoided and consistency is achieved.

IBM CDC also provides a Java class library that can be included in a Kafka consumer application that is intended to consume data delivered by CDC into Kafka. This library, provides:

- Data in the original source log stream ORDER with identifiers available to denote transaction boundaries.
- A mechanism for ensuring exactly one delivery. If there is an interruption in the Kafka environment and data has to be re-sent, a consumer can be developed to only consume and process the data once.
- A bookmark that can be used to restart the consuming application from where it last left off processing.

Also available in CDC delivery are sample consuming applications that show how to poll records that were read by the Kafka transactionally consistent consumer in order to:

- Write them to the standard output in the order of the source operation.
- Publish them in text format to a JMS topic.

Users are free to adapt the samples to suit their needs or to write their own consumer applications.

Feeding Kafka for use as an analytics source of *data in motion*

Given the available processing and storage capabilities available in the Kafka platform, some organizations are choosing to exploit the analytics capabilities of Kafka and the wider system around Kafka, such as the Spark platform. As part of this strategy, IBM data replication can be used to provide real-time feeds of transactional data from mainframes, and distributed environments into Kafka topics so that standard data consumers can deliver the data into the data lake.

Target capabilities

Big data target

Capabilities applicable to both the Kafka and Hadoop targets:

- Hadoop and Kafka target integration with all captures and sources, including those from DB2 z, IMS and VSAM, Oracle, DB2 LUW, Microsoft SQL Server, Informix and Sybase.
- Fixed-price licenses are available for the target apply components.
- Tightly integrated IBM data replication and Hortonworks' Hadoop and Kafka distributions.

Kafka target

Key Kafka target apply capabilities include the following:

- Expansive control of message format and delivered to users' choice of topics or topic partitions in Kafka.
- Reconstruct full consistent transactions in order created with included consumer sample.
- Outstanding performance due to built-in parallelism.
- Source schema information preserved in the target Kafka cluster.
- Kafka topics support predefined formats and custom formats.

Hadoop target

Key Hadoop target apply capabilities include the following:

- Standard WebHDFS interface, utilizing industry standard REST APIs.
- Customized output formats via an included custom formatter and audit trail capabilities to map HDFS file records back to operations on the source database.

Alternatively, Kafka consuming applications can perform analytics functions using the data amassed in the Kafka clustered file system itself or for triggering real time events. For example, when a dormant customer accesses their account, a Kafka consumer application could send a "welcome back" email.

A third common use case is the building of a near real-time audit environment that supports compliance mandates such as GDPR. This is achieved by using the replicated *before-and-after* images that contain information about the source of the changes.

Conclusion

With IBM data replication, every organization can become more agile and improve performance with timely visibility into the data that drives their activities. IBM data replication provides dynamic, near real-time, incremental delivery of transactional data to a broad spectrum of databases and big data targets including Kafka and Hadoop. Access to IBM's expertise, industry leadership and collaboration with other big data domain leaders such as Hortonworks, ensures Hadoop and Kafka distributions are tightly integrated with data replication.

For more information

To learn more about Kafka and Hadoop best practices and IBM data replication solutions, please contact your IBM representative or IBM Business Partner, or visit: **ibm.com/** data-replication/

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
May 2018

Please Recycle

IMS14530-USEN-02