IBM InfoSphere Data Replication for Big Data

Provide near real-time incremental delivery of transactional data

Highlights

- Stream changes in real-time in Apache Hadoop or Kafka data lakes or hubs
- Provide agility to data in data warehouses and data lakes
- Achieve minimum impact on source systems using log-based captures
- Replicate data through support for a wide variety of sources and targets

Organizations are making big improvements to their data architecture and data replication is playing a key role. As big data technology platforms continue to build value for better business insights, the mounting pressure on leaders to enable their entire organization to unlock trusted data assets has led to modernizing solutions for data warehousing. What used to be a single data warehouse node is shifting to a distributed architecture.

Traditionally, having real-time data available to enterprise data hubs or data lakes has often been a challenge. To remain flexible and agile in the big data world, enterprises need to capture information with low impact to source systems, deliver changes to analytics and other systems at low latency and analyze massive amounts of data in motion. The IBM InfoSphere 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 so that organizations can capitalize on emerging opportunities and build a competitive advantage through more real-time analytics. This automation unlocks a multitude of data stores to the business and customer facing platforms that are critical to success in the dynamic and connected environments organizations are building.

IBM Data Replication

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. Including an Apache Hadoop data lake, Apache Kafka landing zone or streaming hub, cloud data store and transformation engines such as IBM InfoSphere Information Server DataStage.

Homogeneous data replication scenarios are also supported, such as version-to-version database migration, maintenance of active and stand-by and active and active data environments for high availability or database mirroring for workload balancing across systems.

IBM Data Replication's Kafka and Hadoop target engines are powered by the standard replication architecture and capabilities: fault tolerance, scalability, security, data lineage and auditing. Additionally, access to IBM expertise and collaboration with big data domain leaders, such as Hortonworks, ensure that clients are surrounded with industry leading support.

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 various enterprise customers 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 applications can continue normal operations while replication runs continuously. IBM Data Replication captures data from database logs but not from triggers or selects from tables. This ensures that the performance of even the most demanding mission-critical applications running on the source system aren't adversely affected. *Figure 1.*

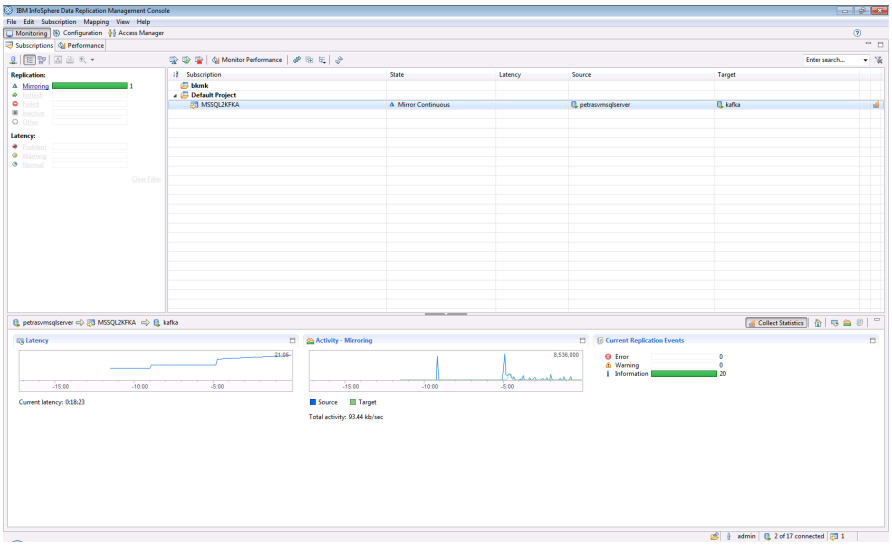

Figure 1: Shows the monitoring dashboard in the Management Console GUI.

Ready for enterprise scale operational demands

In addition to providing standard GUIs for centralized operation and ease of use, recognizing that enterprise customers expect near lights out operations, 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.

Industry standard security

Supporting industry standard security best practices and technologies, security for user access is guaranteed via Lightweight Directory Access Protocol (LDAP) based authentication. 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 essential for preservation of units of work and to maintain referential integrity. The software supports full transaction granularity with available 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 also choose to preserve data lineage and available metadata associated with database changes. Using such capabilities, before-and-after images of record updates or metadata, or both. So that the user who made the change in the original transaction can be replicated and saved on the replication target. *Figure 2.*

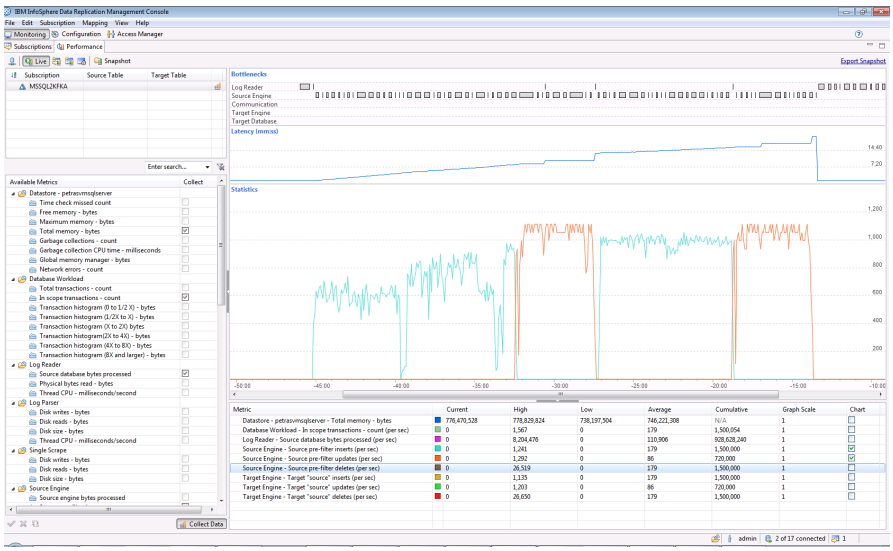

Figure 2: Monitor performance from the Management Console.

Consistent performance and scalability

The IBM Data Replication solution is proven technology that has been successful deployed by a wide variety of enterprise customers who have repeatedly proven out the technology's scalability and performance in the most challenging implementations.

Up-to-date availability of data is often essential for the consumers of your data to make the right decisions at the right time, minimizing latency. With data being delivered in the most efficient and effective way possible, you can deliver incremental changes with very low latency to the target systems for better business agility.

IBM Data Replication addresses all of the above needs while supporting a broad range of operating systems, and all major databases, Kafka, Hadoop and more as replication end points.

Real-time integration using Hadoop and Kafka

As organizations consume and drive insights with big data, merging operational transaction data such as customer, product and accounting data with high-volume data streams such as smart devices, social media and web interactions is necessary. The IBM Data Replication portfolio provides this integration by targeting Hadoop and Kafka. *Figure 3.*

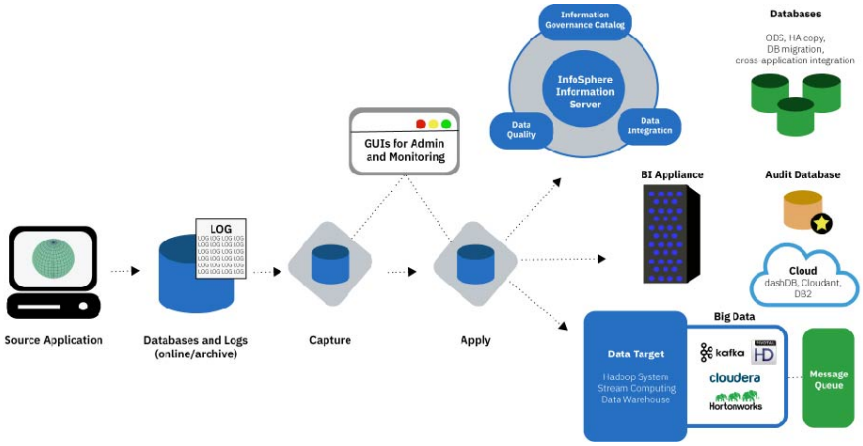

Figure 3: IBM Data Replication log-based Change Data Capture (CDC) with real-time feed to Apache Kafka or Hadoop clusters and many other targets.

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. Hadoop is ideally suited to the storage and analysis of vast amounts of unstructured data.

Built from the ground up for high performing access in service of analytical applications and information exploration, the typical enterprise will use the Hadoop platform to bring together its structured business information with the very high volume and unstructured data from social media, internet activity, sensor data and more. IBM Data Replication supports this strategy by delivering real-time feeds of transactional data from mainframes and distributed environments directly into Hadoop clusters via its Change Data Capture (CDC) Hadoop target engine using a WebHDFS interface.

The analytics software that uses the Hadoop clusters as their system of reference are then able to benefit from the unlocked transactional data assets made available by data replication for 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 must 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 such as, avoiding millions of files is also required. Performance and availability of the HDFS cluster is directly impacted by the maintenance or lack thereof.

Additionally, files in HDFS have fixed formats such as a comma-separated fields. There's 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's 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. It's 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 Java Script 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 like that is 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 in 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, such as a seven day or 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 wanted 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 IBM Db2 Hosted.

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

Given the available processing and storage capabilities available in the Apache Kafka Platform, some organizations are choosing to use the analytics capabilities of Kafka and the wider eco system around Kafka such as the Apache Spark Platform. Again, 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.

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, a Kafka consumer application that subscribes to relevant Kafka topics could send a “welcome back” email to a customer in response to data replicated into Kafka that indicates that a long dormant customer has just accessed their account.

A third common use case is to use the replicated before-and-after images written to Kafka that also contain information about the source of the changes to build a near real-time audit environment in support of compliance mandates such as General Data Protection Regulation (GDPR).

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 IBM Db2 for z/OS, IBM IMS and VSAM, Oracle, Postgre SQL, IBM DB2 for Linux, Unix and Windows, Microsoft SQL Server, Informix and Sybase
- A remote capture source engine for IBM Db2 z/OS sources that provides a second operational model for capturing changes from IBM Db2 z/OS that helps reduce z/OS MIPS needed to replicate IBM Db2 IBM Z data and which could help reduce the dependency on specialized IBM Z skills for deployment and monitoring of replication.
- Fixed priced 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

Hadoop Target

Key Hadoop Target apply capabilities include the following:

- Standard WebHDFS interface, using 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

Why IBM?

IBM supports enterprise DataOps with data integration tools to transform structured and unstructured data and delivers it to any system.

IBM Data Replication solutions provides trusted data integration and synchronization to efficiently manage data growth. It powers the use of real-time information for DataOps by enriching big data systems and mobile applications, even capturing data that is constantly changing.

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.

Learn more about DataOps at ibm.com/dataops. Follow us on Twitter at @IBMData, on our blog at ibmbigdatahub.com and join the conversation #DataOps.

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: ibm.com/financing