Insight Brief

Driving Data Replication Initiatives: A Use Case and Comparison of the Latest Software Tools

VEERENDRA BULUSU, Engagement Manager, IQVIA Global Pricing & Contracting

Table of contents

In today's world of increasing competition, business regulations, and operational costs, data and analytics serve as the lifeblood across all industries and are of paramount importance in making timely enterprise-level decisions. It is vital to keep information synchronized across all business-critical systems, as enterprises look to different technologies and software applications to streamline their business operations and derive valuable insights. As a result, the number of integrated systems within an organization continues to increase, and IT teams now need to develop a data replication strategy to ensure that their data is synchronized correctly across the system landscape.

Data replication involves duplicating and storing identical copies of data in multiple locations, which is essential for disaster recovery, enhancing data access performance, ensuring fault tolerance, and facilitating data backup and archiving. Having a sound data replication strategy enables the accurate and timely reporting required to make the key decisions that safeguard competitive advantages and fuel future growth.

Synchronous vs. asynchronous data replication strategies

Based on an organization's requirements, the strategy to replicate data across systems may be a **synchronous** process in which the data is replicated in near real-time, or an **asynchronous** process that replicates data as intermittent updates typically executed in batches. Strategy selection should take into consideration the necessity for data consistency and permissible latency. Additionally, replicated data requires strong security protocols to avoid unauthorized access or unintended alterations as data is transferred across different cloud instances and processed using different technologies and software programs. Multiple software tools have been developed to satisfy an organization's unique data replication needs, and for this discussion, we will take a look at Oracle GoldenGate, Databricks, and Informatica Intelligent Cloud Services (IICS).

Data replication tool comparison

There are several key data replication criteria to consider when selecting a replication tool.

Comparison Criteria	Oracle GoldenGate	Databricks	Informatica Intelligent Data Management Cloud (IICS)	
			Replication Task	Dynamic Mapping Task
Additional license cost	Yes	No	No	No
Level of effort to implement	Medium	Medium	Low	Low
Replication target	Any target	Any target	Any target	Any target
Data/divisional security	Yes	It requires customization	It requires customization	It requires customization
Real time replication	Yes	No	No	No
Available scripting languages	SQL, PL/SQL, Shell scripting	Python, R, SQL, PL/SQL, Scala, Java, Shell scripting	Python, SQL, PL/SQL Java, Shell scripting etc.	Python, SQL, PL/SQL Java,Shell scripting etc.
Available as a cloud service	Yes	Yes	Yes	Yes
Scheduler capability	Yes	Yes	Yes	Yes
S/W upgrades/patches	Managed by vendor	Managed by vendor	Managed by vendor	Managed by vendor
Replication type	Full/incremental	Full/incremental	Full	Full/incremental
Ease of use	Medium	Medium	Low	Low
Performance	High	High	Medium	Medium

In summary:

- **Oracle GoldenGate** offers real-time replication and broad support, but demands more technical expertise, incurs additional licensing fees, and has restricted platform compatibility.
- **Databricks** provides adaptability well-suited for intricate data processing, though it may result in fluctuating expenses.
- **Informatica Intelligent Data Management Cloud** is focused on the user and oriented towards cloud computing. However, it may face delays, and it lacks incremental replication capabilities for databases.

A data replication use case

The need

A leading pharmaceutical company utilizes an Amazon Web Services (AWS) cloud-based Revenue Management solution to support its contracts and pricing operations, encompassing areas like Master Data Management — Customers & Products, Pricing, Chargebacks, Commercial Rebates, Medicaid, Government Pricing, and utilization script-level data validations. To meet its enterprise reporting needs and still maintain data security control, the manufacturer decided to only replicate data from selected tables from the Revenue Management (RM) Cloud to the data warehouse Lake layer, which is also hosted on AWS.

The solution

A team from the IQVIA Global Pricing & Contracting (GPC) practice was engaged to evaluate the manufacturer's data replication requirements and develop a solution that was also aligned to the overarching enterprise data replication strategy and toolset.

The IQVIA team recommended a novel approach to replicate the data from 80+ Revenue Management system tables by using the Informatica Cloud Data Integration (IICS) — Dynamic Mapping method. This approach is an alternative to more conventional data replication techniques that typically require writing tailored code for every data source and destination, an expensive and time-intensive process.

The team developed a reusable template that engages the IICS Dynamic Mapping feature to ingest data into the data lake. Dynamic Mapping tasks leverage the power of IICS to automatically generate data mappings based on the source and target data schemas. This allows the movement of data from any database table to the data lake with just a few configuration changes.

IICS's reusable template for Dynamic Mapping tasks can be used by any organization looking to move data from a source system's database to any number of systems across the enterprise. The template is agnostic to any source platform, making it a valuable, reusable tool for the client.

The benefits realized

Time efficiency

Accelerates table ingestion by 80% through parallel processing.

Cost reduction

Avoids the need for custom coding for each data source and target, leading to substantial savings in development and upkeep. It supports multiple sources and various targets.

Enhanced flexibility

Companies can swiftly transfer data from any source table to any data lake destination without waiting for IT to develop custom scripts.

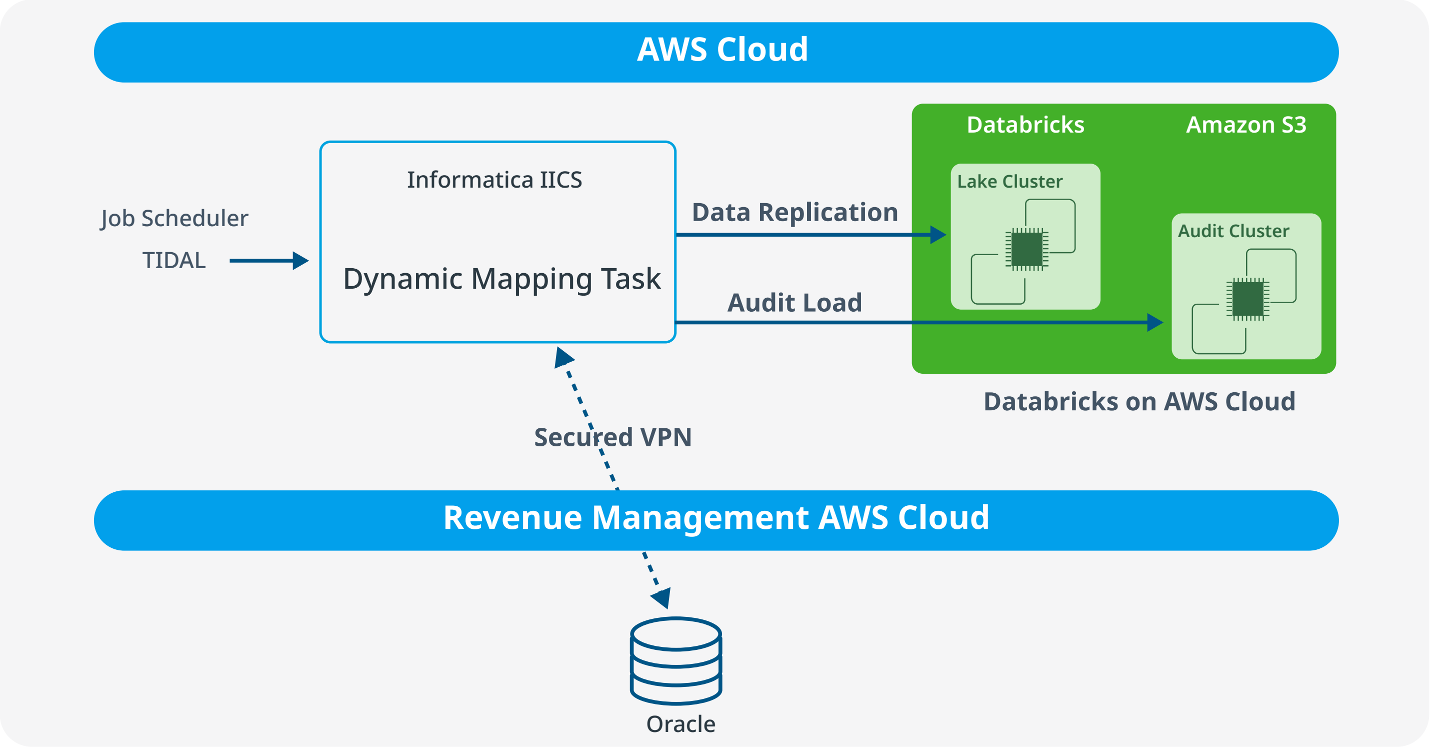
Table selection

Selecting specific tables can reduce resource usage and cut down on ingestion time.

Implementation considerations

This replication is achieved using the IICS Dynamic Mapping approach, which replicates selected tables into Databricks delta tables on AWS. The orchestration of this asynchronous process is managed using the Tidal tool. Below is a high-level view of this process:

High Level Process Flow



Implementation challenges

Below are some challenges faced during the implementation and the strategies adopted by the team to overcome them.

Unmarshall errors

- An 'Unmarshall Error' occurs when a program fails to reconstruct the equivalent memory structure. This error was encountered when the Unity catalog setting was enabled in Databricks. The team resolved the issue by using the schema name as a parameter in IICS.

Oracle to Databricks data type conversion challenges

- From Oracle's RAW to Databricks' Binary Format
  - » In Oracle, the RAW datatype is used to store binary information. For compatibility with Databricks, the RAW datatype has been mapped to the Binary type, which Databricks natively supports.
- Unwanted Additional Zeros in Databricks' Decimal Precision
  - » The team changed the column mappings to DECIMAL with the same precision as the source. However, Databricks added extra "0"s due to its decimal precision.

- Inconsistencies in DATE Data Type Handling
  - To address this issue, the team loaded all Date columns from the source database as Timestamps in Databricks.
- Precision Loss in Timestamp Data Type Column
  - The ‘Last Modified’ column, which has a ‘Timestamp’ data type in both Oracle and Databricks, shows precision up to 6 microseconds. During data transfer from the source to Databricks Delta tables, the last 3 microseconds were lost, leading to discrepancies in data validation due to a limitation in IICS.

Performance bottlenecks

- When dealing with very large data sets, we encountered unexpected delays due to memory limitations in Informatica Intelligent Data Management Cloud (IDMC) within the lower development environments. Adjusting the memory allocation resolved this issue.

Limitations

- Dynamic mapping tasks are not designed for complex data alterations or transformations. If your replication requires extensive data cleaning, filtering, or other modifications, you may need to combine the dynamic mapping task with other data transformation tasks within IICS.
- The error handling and logging features within dynamic mapping tasks may be less detailed compared to specialized data integration tools. Troubleshooting complex issues might necessitate extra effort to identify the underlying problem.

Data reconciliation approach

After successfully replicating the data from the Revenue Management system, the team implemented an automated Data Reconciliation Method using Python/SQL framework. Below is an outline of each step.

Framework for reconciliation:

- A framework is established to compare data between source and target systems. This framework employs Python/SQL to define the comparison logic.

Oracle driver setup:

- To connect to the Oracle database from Databricks, a compatible Oracle driver is installed in the Databricks cluster.

Data comparison:

- The framework, with the driver installed, executes a simple SQL command to compare data, including table structures, field values, and record counts, ensuring they match the source.

Python automation:

- The reconciliation process is automated by parameterizing the tables and columns for comparison using Python.

Reconciliation reporting:

- Post-replication, the reconciliation framework is activated. Any discrepancies between source and target data are documented in a report. These reports are stored in an AWS S3 folder named “QC_Reports” (Quality Control Reports) for easy access and review of data quality issues with the business.

Conclusion

Every company faces unique organizational challenges driven by their enterprise data and reporting needs, as well as their enterprise-level strategies and selected technologies. Despite this uniqueness, one commonality is the need for tailored solutions to drive accurate and synchronized data at an enterprise level. When it comes to data replication, the choice of the software tool depends on the organization's functional and technical requirements, the data replication criteria that need to be met, and the tools already available and/or the budget to purchase new software. With many tools on the market, organizations and IT teams have options, such as IICS Replication Task for user-friendly replication, Databricks for complex transformations, and GoldenGate for real-time performance. Ease of use, transformation capabilities, and real-time performance are all considerations to make an informed decision.

The IQVIA GPC team specializes in automating Data Management and related processes, offering expertise to reduce manual work and to develop innovative, reusable solutions that continue to provide value for the customers that we work with. For more information or assistance, please contact John Wu (john.wu@iqvia.com) to start the conversation.

CONTACT US

John Wu, Global Pricing & Contracting
john.wu@iqvia.com
iqvia.com