ATTUNITY

Accelerate Your Data Pipeline for Data Lake, Streaming and Cloud Architectures

INTRODUCTION
—

Analysis of a wide variety of data is becoming essential in nearly all industries to cost-effectively address analytics use cases such as fraud detection, real-time customer offers, market trend/pricing analysis, social media monitoring and more. The use of machine learning and other artificial intelligence (AI) algorithms, drawing on sources that include Internet of Things (IoT) devices, further increases the required volume, variety and velocity of data.

Integrating this data at scale, however, can overburden IT with complex manual coding (often varying by platform type) and procedures that disrupt production sources. In such environments, many data architects and database administrators struggle to efficiently execute and track replication across the enterprise. They also lack the right tools to efficiently manage the hundreds or potentially thousands of replication tasks these initiatives entail.

In short, modern data requirements break traditional data integration tools.

Attunity changes the landscape with Attunity Replicate, the flagship offering of its modern data integration platform. Now, organizations can efficiently deliver more data, ready for agile analytics, to diverse data lake, streaming and cloud architectures.

Modern Data Integration

Attunity Replicate modernizes your environment by moving data at high speed across all major source and target platforms with a single "Click to Load" interface that completely automates the end-to-end replication process. With this solution, administrators and data architects can easily configure, control and monitor bulk loads and real-time updates with enterprise-class change data capture (CDC) capabilities that enable instantaneous database replication of changed data to the target. Our zero-footprint CDC eliminates any risk of production impact.

Attunity Replicate accelerates both heterogeneous and homogeneous data replication, and controls data flow across hybrid multi-platform environments. Attunity Replicate supports all major databases, including Oracle, SQL Server and DB2. In addition to transactional database support, Attunity Replicate also integrates with all the major analytics platforms, including HPE Vertica, IBM PureData (f.k.a. Netezza), Microsoft PDW, Oracle Exadata and Teradata UDA; Hadoop distributions from Cloudera, Hortonworks and MapR; and streaming systems such as Apache Kafka. Attunity Replicate leverages native utilities and APIs to guarantee fast, optimized and secure data capture and loading for new targets. See a detailed list of supported platforms in the Appendix.

The simplicity and elegant design of Attunity Replicate delivers universal data availability and ease of use of automation.

Ease of use and automation

A key Attunity Replicate differentiator is the "Click-2-Replicate" user interface that enables drag-and-drop functionality. This intuitive approach to replication is easy to learn and fast to implement. Because Attunity Replicate automates the steps required to build a replication solution, it shields users from complexity, eliminating the need for master DBA skills, custom scripting, or consultants. The "Click-2-Replicate" designer is a web-based interface that users can access from anywhere to configure database schema mappings between sources and targets, transformations, and filtering-- all using a graphical task map. The designer helps users create table selection patterns, configure transformations, and define filters easily and rapidly.

Industry's broadest platform support

For years, Attunity has focused on developing deep partnerships and broad product integration with industry leaders. We support all major source and target systems for data replication, including relational database, data warehouse, data lake, Hadoop, cloud and mainframe platforms. Attunity Replicate also supports MongoDB as a NoSQL target and writes change data capture as messages to all major streaming platforms.
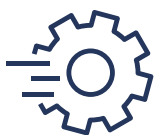
Capabilities and Architecture

Attunity Replicate's multi-server, multi-task, and multi-threaded architecture helps customers scale their on-premises and cloud implementations to thousands of distributed servers and data centers worldwide. The architecture of Attunity Replicate is comprised of three domains: sources (databases, etc.), replication server(s), and targets (databases, data warehouses, Data Lakes, cloud, etc.). End users interact with these domains through the Attunity "Click-2-Replicate" designer and the Attunity Enterprise Manager. Key architectural principles include full load and change data capture replication; agentless, zero software footprint at source and target; scalability and flexibility; the "Click-2-Replicate" user interface, and centralized visibility and control.
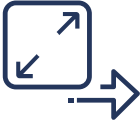
Full load replication

With full-load replication, Attunity Replicate takes all the tables from the source and creates copies at the target. Then, metadata required by the target is automatically defined and the tables are populated with data from the source.

Data is loaded into one or multiple tables to improve efficiency. Although source tables may be subject to update activity during the full load process, there is no need to stop applications in the source. The CDC process automatically activates when table loading starts. However, changes are not applied to the target until after loading is completed. Although data on the target may not be consistent while the load is active, upon its completion the target has full data consistency and integrity.

In addition, the loading process can be interrupted. When it restarts, it will continue from the point it was stopped. New tables can be added to an existing target without reloading existing tables. Similarly, columns in previously populated target tables can be added or dropped without the need to reload.

END TO END PROCESS AUTOMATION

Schema/DDL replication

Attunity Replicate automatically generates target databases based on metadata definitions in the source schema. Any DDL changes made to that schema, such as the addition of new tables and columns or changes to data types, can be replicated dynamically to the target.

Incremental replication or change data capture (CDC)

The Change Data Capture (CDC) process copies updates as they occur in the source data or metadata and applies them to the target endpoint in real time. Attunity Replicate CDC is designed to move large volumes of data changes into target databases or cloud environments with speed, ease and efficiency.

*Attunity Replicate offers the following options for CDC:*

• Log based capture
CDC operates by reading the recovery log file of the source endpoint management system and grouping together entries for each transaction. The process employs techniques that ensure efficiency without impacting the latency of the target data. If the CDC process cannot apply the changes to the target within a reasonable time frame (e.g., when the target is not accessible), then it buffers the changes on the Replication server for as long as necessary. There is no need to reread the source DBMS logs, which may take a long time.

• Query-based capture
Log-based capture is not available for certain data warehouse sources, such as Teradata or Microsoft PDW platforms, since these endpoints don't generate transaction logs. In these cases, Attunity Replicate queries the source tables using context columns, such as TIMESTAMP, to identify and capture changes efficiently from source EDW platforms.

FLEXIBLE CHANGE DATA CAPTURE OPTIONS
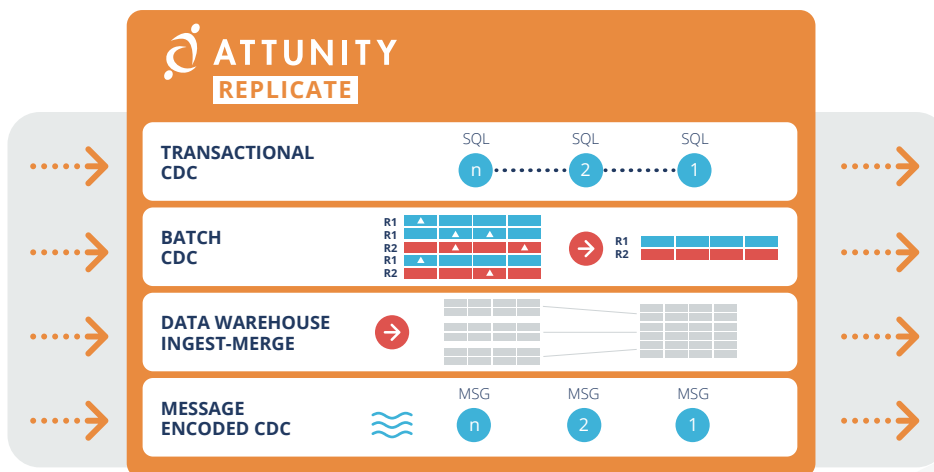
Advanced change data capture technology (CDC) provides several options when delivering data to targets:

• Transactional CDC
Transactional CDC is an option for standard database targets where transactional consistency is more important than high performance. Attunity Replicate streams changes on a transaction-by-transaction basis to maintain the transactional integrity at any point in time.

• Batch CDC
In cases in which high transaction rates and low latency are required, Batch Optimized Apply for CDC, also called simply "Batch CDC," is the right choice. In this case, a pre-processing action occurs to group transactions into batches in the most efficient way.

• Data Warehouse Ingest-Merge
This process follows the same step as Batch CDC, but adds a final step by leveraging the target data warehouses' performance-optimized native utilities as it delivers the data.

• Message Encoded CDC

Attunity Replicate integrates with Apache Kafka and other streaming systems to ingest high data volumes from many data sources, enabling administrators to deliver data to HBase, Cassandra and other Big Data platforms. They can load data either in bulk or through Replicate CDC, which uses Kafka message brokers to relay source changes automatically through in-memory streaming. Attunity Replicate supports multi-topic and multi-partitioned data publication, and as with other platforms, provides integrated management and monitoring through an intuitive console. Attunity Replicate supports additional message streaming targets such as Azure Event Hub and MapR-ES.

Zero software footprint

Data replication must ensure performance and continuous availability, both for production applications and analytics users. Attunity Replicate's unique zero-footprint architecture is designed so that CDC processes can identify and replicate production transactions real-time with a remote transaction log reader. No agents are required on the source or target databases, eliminating administrative and performance overhead on mission-critical systems. This delivers a strong advantage over alternative CDC tools that require intrusive triggers and/or shadow tables, slowing production application and replication performance and creating processing and administrative overhead.

For example, Attunity recently released new agentless end-point for IBM DB2 for z/OS and IBM DB2 for iSeries that deliver significant optimizations to improve performance and reduce the footprint when capturing changes from these platforms. Lab tests demonstrate reductions of 85% in source MSU (million service units), 75% in replication latency, and 95% in loading time.

Attunity Replicate is also designed for maximum flexibility. While the transaction log reader can be installed on the replication server to achieve a zero-footprint impact, it can also be installed on the source database server. As a result, filtering of the source rows can be done on either the source database or replication server, whichever the user prefers.
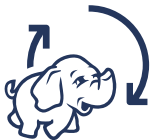
Rapid, automated creation of analytics-ready data stores in Hadoop

Once data is ingested and landed in Hadoop, IT often still struggles to create usable analytics data stores. Traditional methods require Hadoop-savvy ETL programmers to manually code the various steps – including data transformation, the creation of Hive SQL structures, and reconciliation of data insertions, updates and deletions to avoid locking and disrupting users. The administrative burden of ensuring data is accurate and consistent can delay and even kill analytics projects.

Attunity Compose for Hive automates the creation, loading and transformation of enterprise data into Hadoop Hive structures. Our solution fully automates the pipeline of BI ready data into Hive, enabling you to automatically create both Operational Data Stores (ODS) and Historical Data Stores (HDS). And we leverage the latest innovations in Hadoop such as the new ACID Merge SQL capabilities, available today in Apache Hive (part of the Hortonworks 2.6 distribution), to automatically and efficiently process data insertions, updates and deletions. This gives IT and business analysts confidence their data is ACID (Atomicity, Consistency, Isolation and Durability)-compliant.

Attunity Replicate integrates with Attunity Compose for Hive to simplify and accelerate data ingestion, data landing, SQL schema creation, data transformation and ODS and HDS creation/updates.

Time based partitioning

Time based partitioning enables the processing of transactions (insert/update/ delete) from many tables to Hive operational data stores (ODS) or other Hadoop targets in a consistent fashion. Using time-based partitioning, Attunity Replicate will only make available completed transactions for a predefined time interval (minutes, hours or days). Any partial transaction updates will be deferred until the following time. This gives analysts full confidence in the integrity of the data they query and their query results.

Filtering and compression

Whenever filtering conditions are defined on the values of one or more source columns, irrelevant rows and columns are discarded before they are replicated to the target database. This may occur, for example, when a column is not present in the target database schema or when a row does not pass the user-defined predicates on the rows within the source tables.

Transformation

There may be circumstances in which data to be included in the replicated tables is not an exact copy of the source data. Attunity Replicate allows users to define and automatically apply those changes to tables and columns. Examples include renaming the target table, renaming any column in the target table, deleting a target column, changing the data type and/ or the length of any target column, and adding target columns.

Attunity Replicate performs data type transformations as required, calculating the values of computed fields and applying the changes as one transaction to the target. When no user defined transformation is set, but replication is done between heterogeneous databases, some transformation between different database data types may be required. In these cases, Attunity Replicate automatically takes care of the required transformations and computations during the load or CDC execution.

Universal Stream Generation

Databases can now publish events to major streaming services such as Kafka, Confluent, Microsoft Azure Event Hub and MapR Streams. Flexible message formats such as JSON and AVRO, along with the separation of data and metadata into separate topics, allows for smaller data messages and easier integration of metadata into various schema registries.

Optimized Cloud Transfer

Many enterprises have geographically-distributed business units that require data to be stored off premises or in the cloud, in which case each group needs timely access to its own subset of data. To address these needs, Attunity Replicate provides an innovative, highly-secure and resilient WAN transfer engine that optimizes transfer speeds to target databases based on available bandwidth. Algorithms compress large tables and then splits them into multiple, configurable streams. Small tables or CDC streams are batched together to improve efficiency and speed.

Since unpredictable events such as network outages can impact data flow to and from the cloud as well as other remote data repositories, Attunity Replicate offers seamless recovery from interrupted transfers from the exact point of failure. All source data is first staged in files that are located in a temporary target directory. Files are then moved to the target directory, and content between the source and target files is validated. After successful validation, data is loaded into the target database.

Security Capabilities

Attunity has addressed security issues related to data transfer to the cloud by establishing a three-level, secure data transfer mechanism. A secure client-server connection is established through key exchange. An agreed-upon password is then used to scramble the keys, eliminating man-in-the-middle attacks. Files are secured during transfer using advanced, NSA-approved (AES-256) encryption.

Intelligent Management and Control of Enterprise Data Integration

Attunity Enterprise Manager (AEM), an extension of Attunity Replicate, enables efficient, high scale data replication for initiatives such as Data Lake consolidation. AEM provides centralized control of Attunity Replicate tasks and data flow across distributed environments, enabling enterprises to scale easily and monitor thousands of integration tasks in real-time through KPIs and alerts. With AEM, organizations can monitor distributed Replicate servers across multiple data centers and control data flow across distributed environments, on premises and in the cloud, from a single console.

AEM provides an intuitive and logical method of improving efficiency, performance and compliance. Customizable views allow you to define how search results are presented, for example grouping tasks by server, database source or target, by specific application, or even by physical location, you can incorporate the enterprise business logic required by regulatory mandates. Granular searching and filtering capabilities offer actionable insight into data loading and task status. By drilling down from the main console, you can view current task status to identify and remediate issues, thereby meeting performance SLAs.

ATTUNITY ENTERPRISE MANAGER: YOUR GLOBAL COMMAND CENTRE

AEM Analytics

AEM helps users manage operations by visualizing data flow performance and resource utilization trends over configurable time periods. You can track historical KPI and activity trends for tasks and servers and pinpoint performance issues and bottlenecks, ensuring replication SLAs are met. You can also regularly measure tasks and system utilization to augment capacity planning and load-balancing decisions. In addition, AEM collects and shares operational metadata with third party reporting tools for enterprise-wide discovery and reporting.
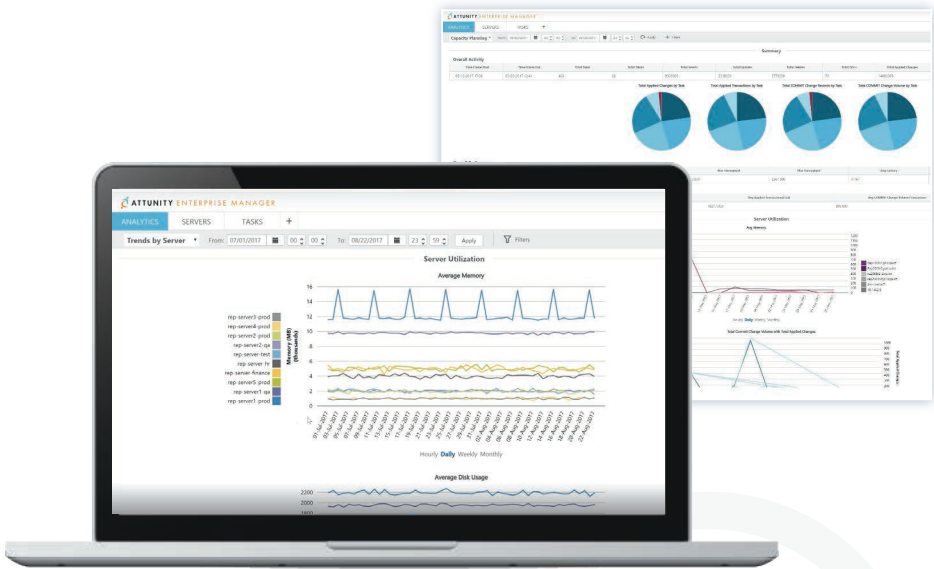
Comprehensive monitoring metrics include:

- ☐ **Change processing throughput graph.** Presents the number of change events or KBs read during a specified time period.

- ☐ **Latency gauge.** Measures the delay between the time a change is visible to the source database and committed, and the time the same change is visible to the target database.

- ☐ **Change processing status circle graph.** Displays information about processed changes, number of insert operations processed, number of update operations completed, and number of metadata or DDL changes processed. Examples of DDL changes include information about events, such as modifications to tables or column names.

- ☐ **Transaction status circle graph.** Presents information about processed changes, commits, rollbacks, and transactions in process.

- ☐ **Notifications and alerts monitoring table.** Displays various alerts and messages about the state of the systems and tasks. Error messages describe processing issues.

Users can start, stop, and reset monitoring through the Attunity Replicate web-based interface from any location at any time.

Microservices API

AEM enables intelligent management and control of hybrid cloud and on-premise deployments at any scale with support for analytics and microservices. New REST and .NET APIs are designed for invoking and managing Attunity services using a standard web-based UI.

Use Cases and Best Practices

Attunity Replicate improves efficiencies and business operations for a variety of enterprise use cases. Typical use cases include data replication for operational analytics or query offloading, high scale Data Lake consolidation, line of business workload offloading, and enablement of cloud analytics.

Enterprise data replication

Using Attunity Replicate as a unified replication platform reduces the time and complexity involved in maintaining data availability across heterogeneous and distributed environments. Attunity Replicate replicates, synchronizes, distributes, consolidates and ingests data across all major databases, data warehouses and Hadoop, on-premises and in the Cloud. With Attunity Replicate, organizations can scale their architectures to move data across thousands of databases with centralized monitoring and management. They are better able to optimize workloads and support business operations, applications, and analytics needs.

Real-Time Data Warehousing

Attunity Replicate enables real-time data warehousing and analytics with change data capture (CDC) technology that delivers database updates to the data warehouse with low latency. With this continuous and efficient data acquisition process, organizations can ensure real-time data availability and eliminate the complexity, cost and delays associated with homegrown and traditional ETL software. Attunity Replicate log-based CDC also minimizes impact on source production operations.

Query Offload and Live Reporting

Organizations use Attunity Replicate CDC to create live replicas from production applications for separate reporting and analytics databases, enabling organizations to scale real-time BI and analytics with zero production impact. By offloading queries and workloads in this fashion, they can continuously analyze fresh data while meeting production application SLAs and minimizing cost.
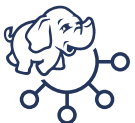
Easy ingestion of structured data into Hadoop

Attunity Replicate delivers high-performance data loading and publication to Data Lake ecosystems through native APIs, and is certified with all the leading Hadoop distributions including Cloudera, Hortonworks and MapR. You also can automatically generate target schemas in HCatalog.

Accelerated Cloud Migrations and Analytics

Attunity Replicate and Attunity CloudBeam (the version of Attunity Replicate that is available on cloud marketplaces) accelerates data flow to the cloud. Organizations can easily and securely transfer data over Wide Area Networks (WANs) at high performance with encrypted multi-pathing technology to all the major public cloud platforms, including Amazon Web Services (all RDS databases and Amazon Redshift), Microsoft Azure SQL Database and Warehouse, and Google Cloud.

SAP Analytics Enablement

Attunity Replicate for SAP easily and securely transfers SAP data, documents, and transactions into Hadoop and other external platforms, and automatically incorporates SAP text and descriptions into target table and column names. Replicate supports data collection from all core SAP business applications such as ERP, HR, CRM, industry-specific applications, and loading data into SAP HANA.

"Click to stream" Big Data with universal streaming integration

Attunity Replicate integrates with Apache Kafka and all other major streaming platforms to ingest high data volumes at low latency from many data sources. This direct integration enables you to feed Hadoop Data Lakes and NoSQL databases such as HBase and Cassandra.

Conclusion

Attunity Replicate is changing data integration by enabling IT to deliver more data, ready for analytics, to data lake, streaming and cloud architectures. Unlike the traditional batch-oriented and inflexible ETL approaches of the last decade, Attunity provides the modern, real-time architecture enterprises require to harness the agility and efficiencies of new data lakes and cloud offerings. Companies seeking a single solution to drive agile analytics and enable universal replication should consider Attunity Replicate.

ATTUNITY REPLICATE
PLATFORM SUPPORT

SOURCES

RDBMS	DW	HADOOP	MAINFRAME
Oracle	Exadata	Hortonworks	DB2 for z/OS
SQL Server	Teradata	Cloudera	IMS/DB
DB2 iSeries	Netezza	MapR	VSAM
DB2 z/OS	Vertica		
DB2 LUW	Pivotal		
MySQL			
PostgeSQL			
Sybase ASE			
Informix			

SAP	CLOUD	OTHER LEGACY	FLAT FILES
ECC on Oracle	AWS RDS	SQL/MP	Delimited
ECC on SQL	Amazon Aurora *	Enscribe	(e.g., CSV, TSV)
ECC on DB2	Salesforce	RMS	

TARGETS

RDBMS	DW	CLOUD	HADOOP
Oracle	Microsoft PDW	Amazon RDS	Hortonworks
SQL Server	Exadata	Amazon Redshift	Cloudera
DB2 LUW	Teradata	Amazon EMR	MapR
MySQL	Netezza	Amazon S3	Amazon EMR
PostgreSQL	Vertica	Amazon Aurora	HDInsight *
Sybase ASE	Sybase IQ	Google Cloud SQL	
Informix	Amazon Redshift	Azure SQL DW	
	Actian Vector	Azure SQL DB	
	SAP HANA	Snowflake	

STREAMING	NOSQL	SAP	FLAT FILES
Azure Event Hubs *	MongoDB	HANA	Delimited
MapR-ES *			(e.g., CSV, TSV)
Kafka			

* Preview release