striim

REFERENCE ARCHITECTURE

Oracle to Databricks

Table of Contents

Overview

This reference architecture is built to assist organizations in migrating and replicating data to the cloud. It covers a quick background on Striim, the infrastructure prerequisites to use Striim, setup process for the source system, the process for writing to the target system, and an implementation guide for Striim itself.

This document is a supplement to Striim documentation located **here**.

If there are any questions, please refer to the Striim documentation.

Background

There's no question about the benefits of cloud computing and public clouds. Rather than maintaining costly internal infrastructure, public clouds enable nimble development and on-demand usage of resources.

Cloud-first strategies where all new development is done on the cloud is now the norm, not the exception. However, existing data locked in on-premises systems remains critical to enterprise operations. This begs the question: How can this on-premises data be liberated to the cloud without affecting any of the existing, mission-critical infrastructure? The answer is Striim.

Striim is a streaming ETL platform that enables real-time continuous Smart Data Pipelines into critical cloud services, empowering migrations of high-value business-critical workloads to the cloud with minimal downtime and risk. Brought to you by the core team behind GoldenGate Software (acquired by Oracle), Striim accelerates data movement into the cloud as a strategic partner of the public clouds.

As seen in the implementation guide below, Striim is a platform which enables a few common use cases:

- **Cloud adoption**, including zero downtime or minimal downtime database migration and database replication from on-premises or other clouds to the cloud.
- **Hybrid cloud data integration**, which typically involves mission-critical systems that will not sunset anytime soon, but need to be continuously replicated to the cloud to take advantage of cloud data services. With Striim, the source and target systems can be kept in sync for years at a time. This is the use case we'll focus on in this reference architecture.
- **Digital transformation**, including real-time data distribution, real-time reporting, real-time analytics, stream processing, operational monitoring, and machine learning. Striim enables Smart Data Pipelines through a complex event processing engine, which allows SQL queries and Java code to be run on the data in flight.
- **Reverse ETL**, where Striim both ingests data in real time into a cloud data warehouse and synchronizes transformed data back to the source systems.

These use cases are possible due to Striim's log-based change data capture (CDC) technology, which reads from the underlying transaction logs of transactional databases. CDC is the least intrusive method of reading from a database, and continuously captures all of the events (Inserts, Updates, and Deletes) that occur on the database itself. Striim has the world's fastest CDC which enables real-time workloads not possible with other technologies.

Setup Considerations

Striim can either run as a fully managed SaaS Striim Cloud, or user managed Striim Platform, typically deployed in the same public cloud as the target data service. Striim Cloud eliminates the additional setup required with Striim Platform, including creating a Linux VM within the target public cloud, installing Java, and i**nstalling the Striim software**. wRegardless of the deployment method, once installed this reference architecture applies to both methods.

Solution

Source Oracle Database

Striim reads from a source system via log-based CDC. Each database engine has a slightly different method of exposing the transaction logs, this reference architecture is focused on Oracle.

Networking

To read from a source database, Striim needs network connection over the typical JDBC ports. In the case of Oracle ensure the default 1521 port is open to the Striim instance, alternatively allow connection over the custom port configured.

Method for Reading

With Oracle, Striim supports three different methods via CDC:

- **Oracle Reader**
  This approach uses Oracle's LogMiner API to read from Oracle's redo logs. This is the default method of reading and does not incur an additional cost.
- **OJet** (Supports 21c, and higher performance)
  OJet is typically used for any use case when reading from Oracle 21c, or for the largest workloads generating large amounts of redo logs.
- **GoldenGate Trail Files**
  Striim also supports reading from GoldenGate trail files. This method is perfect when GoldenGate is already in place in the architecture, and eliminates the need to add another CDC process on the source database.

Performance Impact

CDC, specifically log-based CDC, is the least intrusive method of reading from a database. It is environment-specific as to how much impact there is, but is typically in the low single-digit percentage increase of additional processing power required.

Target Databricks Delta Lake

Prior to building Striim pipelines, there are some prerequisites for the target system.

Schema Creation

Prior to starting the data movement pipelines outlined within the Striim Guide, it is a prerequisite to create the tables in the target system corresponding to those in the source. This can be done using any other tool you prefer, or within Striim via two methods:

- A command line based Schema **Conversion Utility** which creates SQL queries to run on the target system
- Automatically within the UI via the wizard based data pipeline development

Networking

Similar to the source, Striim needs access to the target system. With Databricks, this doesn't require much additional configuration. Striim connects over the typical HTTPs port 443, which should already be enabled.

Permissions

Striim authenticates with Databricks through a Personal Access Token. Please reference **this document** for the required permissions associated with the token.

Method for Writing

Striim currently writes to Databricks through a micro-batch approach. Based off of the user-defined upload policy, events are staged to the DBFS instance specified within the connection properties and written to Databricks using JDBC.

Typically with Databricks use cases it makes sense to only use a time-based interval for the upload policy in order to ensure the upload SLA is consistent throughout the day.

Depending on the use case there is a trade-off between more up-to-date data with a shorter upload policy vs. a larger upload lag with a longer upload policy. A shorter upload policy will incur more Databricks costs, but enable more real-time analytics. Please work with your Striim and Databricks account teams to determine what is best for you.

Smart Data Pipelines

When using Striim with Databricks, there are two modes to run in:

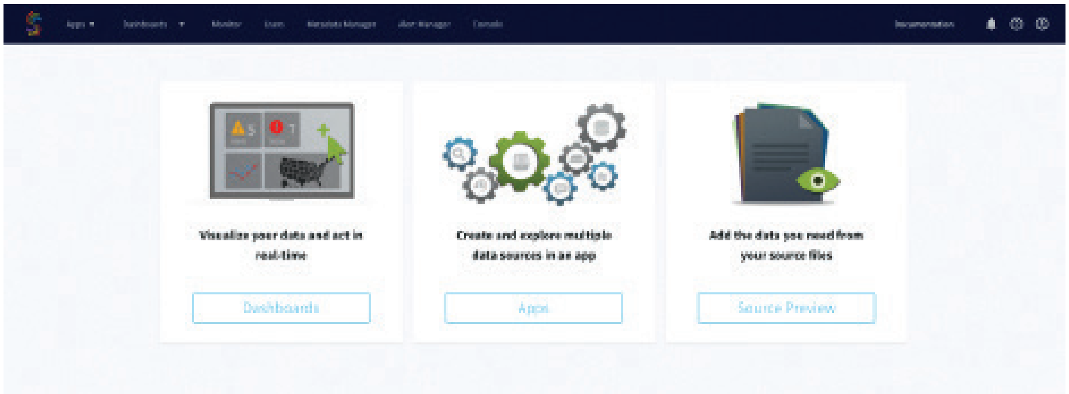
- **Append Only:** Striim writes all events as Inserts, typically used as an audit table to keep track of all of the operations that occurred on the source, then batch jobs are run on after the fact
- **Merge:** Striim will automatically run merge queries on the data warehouse to apply any Update and Delete operations and create an exact replica of the source table on the target

With append only, it is typical to use Striim's in-memory processing capabilities to create Smart Data Pipelines and add an OpType and Timestamp column within the tables to perform batch merge operations after the fact.

Striim Guide

Application Creation – Wizards vs. Drag-and-Drop vs. TQL

Once you've deployed Striim, you'll be greeted by the home page. There are a few different tabs here, but the primary one is applications, which are all of the data pipelines from a source to a target.



Visualize your data and act in real-time

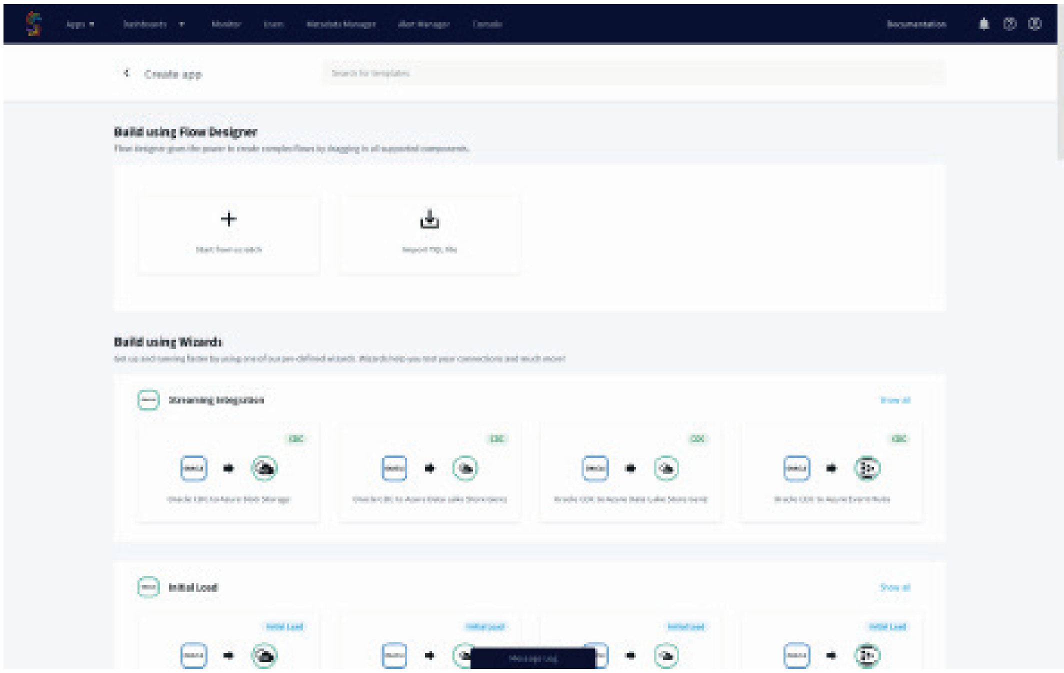
Create and explore multiple data sources in an app

Add the data you need from your source files

Clicking on Apps > Create App will take you to the "Create Application Page" where there are three options:

- Start From Scratch
  This is the main page which uses a drag-and-drop UI and out-of-the-box components to create the applications.
- Import TQL File
  All applications within the UI are backed by what is called a TQL file, which looks like PL-SQL and is a textual representation of the application. These TQL files can be created in a text editor outside of Striim, or programmatically via a simple script, stored in a CI-CD pipeline, and utilize the REST API around Striim to create a programmatic approach around Striim applications.
- Build Using Wizards
  These are out-of-the-box templates which automatically create the application for you.

Depending on the use case one may make more sense than the other. Typically starting with the Wizards or drag-and-drop interface is easiest.



Application Development Process

Most Striim use cases involve three processes:

1. The one-time schema creation process on the target system. This can be done either within Striim with the Wizards or schema creation tool, or outside using another utility.
2. The one-time historical bulk load to seed the target system. When reading from a transactional database, create a Striim initial load application via any approach using the Database Reader and the associated writer. When writing to data warehouses this application should use the load method and large batches to avoid any quota limitations and decrease cost.
3. The log-based CDC application using the appropriate CDC reader and target component writer.

To coordinate the processes and ensure zero data loss, follow the steps found in Striim documentation **here**.

Best Practices

Following the application development process, it is recommended to start with a small subset of the dataset and execute the following steps:

1. Migrate the schema either within Striim or via an outside tool
2. Create a small test initial load application to validate all of the connections. If using the drag and drop interface, start with a simple source component, test the application, then and only then proceed with adding a target. This way can validate that any errors are with the source component.
3. Create a small test CDC application to validate CDC works. Again, start with the source before continuing to add the target component.

4. Implement the zero-data-loss process outlined in the application development process.

5. Proceed to implement this process with a larger data set.

Security

The security process is dependent on the deployment method of Striim:

- Striim Cloud
  Striim Cloud is a fully managed SaaS offering which is managed by Striim. Striim Cloud is SOC2 certified, but not yet HIPAA certified (as of October 2022). Striim Cloud supports **connecting to source or target** systems via a jump server.

- Striim Platform
  Striim Platform is fully run in the end environment, simply moving data from point A to point B within the environment. This typically passes many security questionnaires, as is dependent on the environment's existing security and network protocols.

Summary

After following the above guidance, you should have a fully functional, continuous replication pipeline to the cloud, enabling a continuous 24/7 live feed of data into Snowflake. From there, take advantage of Snowflake's broader ecosystem to analyze and visualize real time data or to make use of machine learning algorithms.

Learn More

Please visit **www.striim.com** or reach out to your Snowflake representative to learn more about other use cases or to ask any questions.

Support

Striim offers additional training, support, and services for more sophisticated deployments than described in this reference architecture. Please reference those before embarking on a different use case.

Contact Us

Please email **partners@striim.com** with any questions, we will gladly schedule a call to discuss your use case in more detail.

About Striim

Striim was founded with a simple goal of helping companies make data useful the instant it's born.

Striim's unified, real-time data streaming and integration platform for analytics and operations collects data in real time from enterprise databases (using non-intrusive change data capture), log files, messaging systems, and sensors, and delivers it to virtually any target on-premises or in the cloud with sub-second latency enabling real-time operations and analytics.

Try it now at **go2.striim.com/free-trial**

Contact us at:

Tel: +1 650 241 0680
Web: **www.striim.com**

striim