Treehouse SOFTWARE

White Paper

Mainframe-to-Cloud – The “Data First” Approach

by
Joseph Brady
Director of Business Development and Cloud Alliance Lead
Treehouse Software, Inc.

Contents

Summary

Since its establishment in 1982, Treehouse Software has garnered an excellent reputation as a leader in mainframe systems software development and marketing, with hundreds of customers in the U.S. and many other countries. Today, Treehouse Software is a global leader in providing solutions for real-time data replication between a variety of mainframe data sources and Cloud and open systems targets, including:

Mainframe sources and targets (for bi-directional replication) include:

- **VSAM**
- **IMS**
- **Db2**
- **CA Datacom**
- **Adabas**
- **CA IDMS**

Cloud Targets are numerous and include:

- **AWS RDS and RDS Aurora**
- **AWS S3 (JSON, CSV, and AVRO)**
- **AWS Kinesis**
- **Kafka (On premises, Confluent and other Clouds)**
- **Azure BLOB Storage (JSON, CSV and AVRO)**
- **Azure SQL Database**
- **Azure SQL Database for PostgreSQL**
- **Azure SQL Database for mySQL**
- **GCP Buckets (JSON, CSV, AVRO)**
- **GCP Cloud SQL: for PostgreSQL, MySQL and SQL Server**
- **On premises and Cloud (SQL Server, Oracle, MySQL, MariaDB and PostgreSQL)**
- **Complete list:** https://treehouse.com/tcVISION_Supported_Environments.shtml

Treehouse Software is a Technology Partner with AWS, Google Cloud, and Microsoft. This white paper highlights our knowledge and insights for implementing successful Cloud and hybrid Cloud mainframe data replication on AWS (Amazon Web Services), Microsoft Azure, and Google Cloud Platform (GCP) using a data replication tool and deep mainframe knowledge. We will outline the phases of a data replication project, including proof of concept (POC), scope, business use cases, resource planning, and deployment.

Data First Approach to Mainframe Modernization

This white paper serves as a guide for organizations planning to replicate their mainframe data on a Cloud platform. Much of an enterprise's mission critical mainframe data is stored in legacy mainframe databases, and the cost to maintain these databases is high. An added complication is that the data is utilized by many interlinked and dependent programs that have been in place for many years, and sometimes, decades. Unlocking the value of this legacy data is difficult due to many very different types of mainframe databases.

Many organizations are now looking for modernization solutions that allow their legacy mainframe environments to continue, while replicating data in real time on highly available Cloud-based platforms. With a "data-first" approach, immediate data replication to the Cloud is enabling government, healthcare, supply chain, financial, and a variety of public service organizations to meet spikes in demand for vital information, especially in times of crisis.

Whether an enterprise wants to take advantage of the latest Cloud technologies, such as analytics, artificial intelligence (AI), scalable storage, security, high availability, etc., or move data to a variety of newer databases, the transition doesn't have to be a sudden big bang. With the correct tools and mainframe knowledge, data synchronization of changes on either platform can be reflected on the other platform (e.g., a change to a PostgreSQL table is reflected back on mainframe). The customer can then modernize their application on the Cloud **without disrupting the existing critical work on the legacy system.**

Benefits

Modernization Implementation speed – Data replication projects can be implemented in a couple of months. This includes the proof of concept and design/architecture stages. After these stages are complete, your organization can start the first production implementation sprint, immediately providing business value. Successive agile sprints allow for incremental deployment of additional file replication, sprint by sprint.

Legacy work can continue while simultaneously benefiting from modern Cloud Services – A data replication product should allow data to be replicated to numerous RDBMS databases, Kafka, JSON, CSV, and AVRO targets allowing mainframe data to be consumed by various Cloud services. These services include machine learning, data warehouses, streaming services, search services, modern programming languages, container services, and more. All this can be accomplished without disruption to the legacy operations.

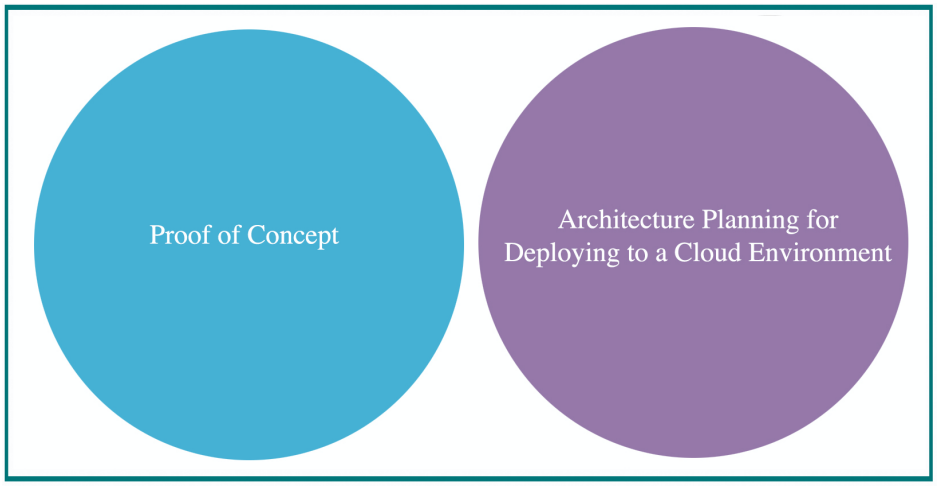
Minimal mainframe programmer requirement – A data replication product should unitize a GUI interface for non-mainframe programmers. Mainframe experts are required in the design/architecture phase and occasionally during implementation. However, the requirement for their involvement should be limited.

Low requirement for mainframe subject matter experts – Application modernization projects require significant legacy subject matter experts, especially for testing. This is generally not the case for data replication projects. Some projects require almost no subject matter expert involvement due to their more technical nature. Some limited

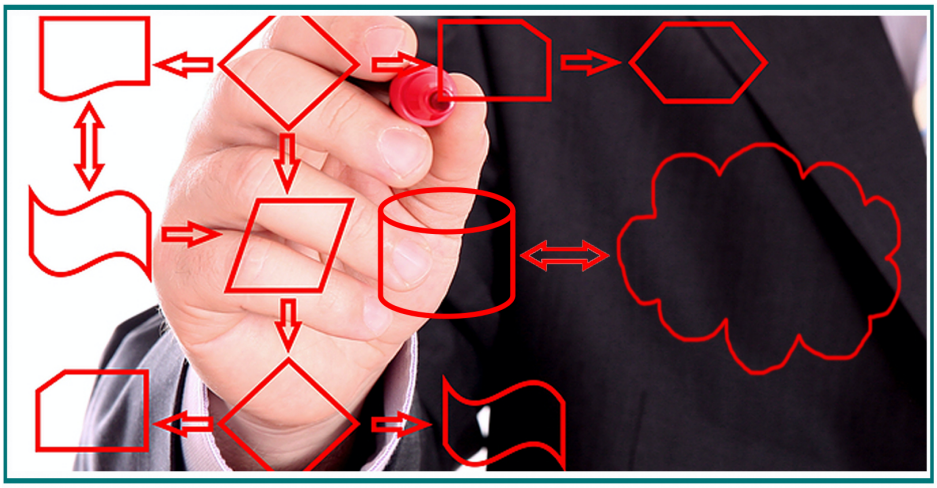
involvement may be required for activities such as identifying personally identifiable information (PII) data, data cleansing, and mapping source-to-target for bi-directional replication.

Minimal part-time project manager required – Often larger application modernization projects require a full-time project manager. Data replication projects require a very part-time project management involvement. This involvement is higher during the design/architecture phases and generally tapers down to a couple of hours per week during implementation.

Treehouse Software considers two phases to be vital to a successful Mainframe-to-Cloud data replication project:



Phase 1: Proof of Concept



Planning of the customer's project requires the identification of 1) Scope and 2) Use Cases. Examples are listed below.

General

The purpose of the POC is to determine fitness for purpose for the ELT/CDC tools. It is important to completely understand the intended use of the product. The first step would be to hold a meeting with key stakeholders to determine the intended use and scope for the product at a high level. Subsequent steps can drill down into the details.

Determine and Document the POC Scope

- The customer must plan and list specific steps for the scope of their data replication scenario. Listed below are some scope considerations for the POC. It is important to understand time constrains and prioritize the scope items based on importance. It may be necessary to push out some lower priority items to meet require timelines. Determine the source databases (typically mainframe databases) that are in scope (e.g., Adabas, VSAM, Db2, etc.)
- Are you testing an on-premises solution, Cloud based or hybrid?
- What are your networking requirements between on premises and Cloud? Will a VPN suffice, or do you need a private connection such as AWS Direct Connect? What are the privacy requirements? Can Cloud services be accessed over the internet, or need a private endpoint such as AWS PrivateLink?
- Determine the database or other targets. Targets can be on-premises bases or Cloud based. Included and be AWS RDS, S3, Azure Buckets, Kafka, PostgreSQL, MySQL, and more.
- Determine the platform requirements. Where will the data replication product administrator component reside? On premises, or in a Cloud VM instance?
- Determine data security requirements. What data will be used? Are there PII requirements? Does data need to be masked? What are the data at rest and data in transit security requirements?
- Do you have any unique requirements (e.g., triggers on source or target)?
- Will you need uni-directional replication or bi-directional replication?
- Will you need to test field addition, deletion?

- Do you have any specific performance requirements? Can they be satisfied with a technical discussion, or are specific proof points required?
- Who are the stakeholders? What is their involvement? Ensure all critical stakeholders are involved to the extent required and determine if they will be involved in the POC or signoff based on specific results.
- What is your product budget? Give your specific component requirements (e.g., Mainframe model, required target database) to the vendor so they can give you budgetary estimate to ensure you are in alignment.

Use Cases

The customer should identify and list their desired use cases for the project. Listed below are some general examples of use cases for a Mainframe-to-Cloud data replication project. Of course, each project will have unique environments and goals. It's important the specific use cases be determined and documented prior to the start of the POC. This will allow the vendor and customer to develop a more accurate POC timeline and have the required resource available at the current times. A general principle should be to prove out the technology using use cases:

- What is the mainframe source database?
- What are the critical issues you need to test? Are there any areas that you believe that would be challenging for the vendor? Examples would be specific transformations, data types (e.g., BLOB), required CDC SLAs, field/column changes, specific security requirement, data volume requirements, etc. Document all of the critical test items.
- Select the minimal set of files (generally 3-10) that will enable you to test all of your critical test items. If there are multiple source databases, ensure specific test use cases are defined for each source.
- What are the target databases? Will you be replicating to a manager Cloud database such as AWS RDS. Are there additional requirements to replicate to S3, Azure BLOB, GCP Cloud Storage, Kinesis, and Kafka? What needs to be tested.
- What are your bulk or initial load requirements? A data replication product should load data directly from mainframe databases or mainframe unloads or image copies. What are the data volumes? Do you have sufficient bandwidth between your mainframe manager and on-premises or Cloud VM to handle your volume requirements?
- What are you Change Data Capture (CDC) requirements?
- Do you have plans for bi-directional replication in the future. What are your specific requirements? Bi-directional replication can be complex and greatly lengthen the POC. Often the customer will perform one bi-directional use case for conceptual proof. Will this suffice for your organization?
- What are your specific high availability requirements? Can they be handled by a technical discussion, or is a specific use case required?
- What are your general security requirements for data at rest and data in transit? Do you have any specific security regulations to follow such as HIPPA or FIPS? What are your PII / data masking requirements. Do you have masked data for the POC or any other specific security requirements for the POC?
- What are your schema requirements? A data replication product should create a default schema based on your input mainframe data. Major changes to the default schema will generally require a staging database.
- Do you have staff available to perform the POC tasks? POCs generally run from 2-4 weeks and you will need part-time staff, 2-4 hours per day. A part-time

mainframe administrator will generally require 2-8 elapsed hours. Other staff will include Windows/Linux/Cloud administrators. 2-4 hours of project management may also be required.

- Are business data transformations required for the POC? A data replication product should handle minor transformation via point and click (e.g., date format transformations). Major transformations that require C++ or product scripting need to be limited to keep within the POC timeframe constraints.
- Are there any triggers or stored procedures? A data replication product should perform CDC replication processing using a database that utilizes these database features.

POC Process

Treehouse Software follows a POC process where our team and the customer agree on use cases and success criteria. This usually starts out with one or more virtual meetings and email exchanges, discussing the number of use cases, databases for source and target, and an estimated schedule. It generally also includes customer success criteria for each evaluation item (e.g., expected replication time).

Treehouse sends out documents regarding POC requirements (e.g., number of open ports required on the mainframe); a POC license agreement for the customer to sign; and an email with the location of the trial software and an associated key.

The customer and Treehouse schedule and execute the POC kickoff (currently by virtual meeting). Included in the meeting are all of the key stakeholders, including all leadership technical staff that will be involved in the POC.

The POC use cases are executed according to the estimated schedule (e.g., Tuesday to Friday 3:00 PM to 5:00 PM for 3 weeks). If technical issues are encountered, Treehouse will work with the customer to resolve them accordingly.

The customer records the results of each use case and records them according to the required criteria.

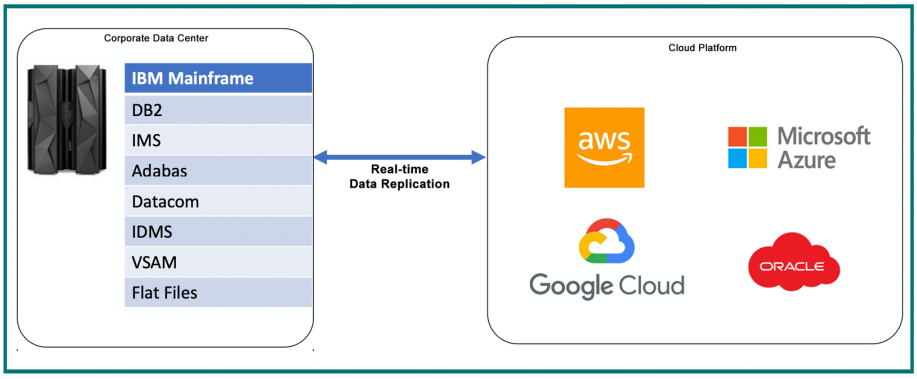
When all use cases are complete, the customer meets with the required staff to determine if they would like to move forward with licensing the product for their replication needs.

After the customer is satisfied with the scope and use cases, discussion points should be identified and listed, for example:

- Identify persistence/data security at each step.
- Identify the target table name in the default schema.
- Determine ability to synch with one or many targets.

A successful, repeatable replication process (with documented results) will be accomplished between the mainframe-based data source and a corresponding Cloud target within the allotted time period.

Phase 2: Architecture Planning for Deploying to a Cloud Environment



A successful move to the Cloud requires a number of considerations and solutions in order to modernize an application on the Cloud. Some examples of these considerations and solutions include:

Personnel Resource Considerations

Staffing for Mainframe-to-Cloud data replication projects depends on the scale and requirements of your replication project. Very large and complex (e.g., bi-directional replication) require more staffing. Staffing can be onshore, offshore, or a combination of on-shore / off-shore.

Most customers deploy a data replication product with Windows and Linux knowledgeable staff with varying levels of seniority. For the architecture and setup tasks, we recommend senior technical staff to deal with complex requirements around the mainframe, Cloud architecture, networking, security, complex data requirements, and high availability. Less senior staff are effective for the more repeatable deployment tasks such as mapping new database/file deployments. Business staff and system staff are rarely required but can be necessary for more complex deployment tasks. For example, bi-directional replication requires matching keys on both platforms and their input might be required. Other activities would be PII consideration, specifics of data transformation and data verification requirements.

An example of staffing for a very large deployment might be one very part-time project manager, a part-time mainframe DBA/systems programmer, 1-2 staff to setup and deployment the environment and an additional 1-2 staff to manage the existing replication processes.

Environment Considerations

As part of the architecture planning, your team needs to decide how many tiers of deployment are needed for your replication project. Much like with applications you may want a Dev, QA and Prod tier. For each of these tiers, you will need to decide the level of separation. For example, you might combine Dev and QA and not Prod. Many customers will keep production as a distinct environment. Each distinct environment will have its own set of resources, including mainframe managers (possibly on separate LPARs), Could VMs (e.g., EC2) for replication processing and managed Cloud RDBMS (such as AWS RDS).

After the required QA testing, changes are deployed to the production environment. Object promotion test procedures should be detailed and documented, allowing for less experience personnel to work in some of test tasks. Adherence to details, processes, and extended testing is most import when deploying replication with bi-directional replication due to the high impact of errors and difficult remediation.

Rollout Planning

A data replication product is typically deployed using Agile methods with sprints. This allows for incrementally realized business value. The first phase is typically a planning/architecture phase during which the technical architecture and deployment process are defined. Files for replication are deployed in groups during sprint planning. Initial sprint deployments might be low value file replications to shield the business for any interruptions due to process issues. Once the team is satisfied that the process is effective, replication is working correctly, and data is verified on the source and targets, wide scale deployments can start. The number of files to deploy in a sprint will depend on the customer's requirements. An example would be to deploy 20 mainframe files per 2–3-week sprint. Technical personnel and business users need to work together to determine which files and deployment order will have the greatest business benefit.

Security

For security, both on-premises and to the major Cloud environments, there are several considerations:

- Data will be replicated between a source and target. The data security for PII data must be considered. In addition, rules such as HIPPA, FIPS, etc. will govern specific security requirements.
- The path of the data must be considered, whether it is a private path, or if the data transverses the internet. For example, when going from on-premises to the Cloud the major Cloud providers have a VPN option which encrypts data going over the internet. More secure options are also available, such as AWS Direct Connect and Azure Express Route. With these options, the on-premises network is connected directly to the Cloud provider edge location via a telecom provider, and the data goes over a private route rather than the internet.
- Additionally, Cloud services such as S3, Azure Blob Storage, and GCP buckets default to route service connections over the internet. Creating a private end point (e.g., https://docs.aws.amazon.com/AmazonS3/latest/userguide/privatelink-interface-endpoints.html) allows for a private network connection within the Cloud provider's network. Private connections that do not traverse the Internet provide better security and privacy.
- Protecting data at rest is important for both the source and target environments. The modern Z/OS mainframe has advanced pervasive and encryption capabilities: https://www.redbooks.ibm.com/redbooks/pdfs/sg248410.pdf. The major Cloud providers all provide extensive at-rest encryption capabilities. Turning on encryption for Cloud Storage and databases is often just a parameter setting and the Cloud provider takes care of the encryption, keys, and certificates automatically.

- Protecting data in transit is equally important. There are often multiple transit points to encrypt and protect. First, is the transit from the mainframe to on-premises to the Cloud VM instance. A mainframe data replication product should provide protection employing TLS 1.2 to utilize keys and certificates on both the mainframe and Cloud. Second is from the Cloud VM to the Cloud target database or service. Encryption may be less important since often these services are in a private environment. However, encryption can be achieved as required.

High Availability

- During CDC processing, high availability must be maintained in the Cloud environment. The data replication product should keep track of processing position. The first can be a Restart file, which keeps track of mainframe log position, target processing position, and uncommitted transactions. The second can be a container stored on Linux or Windows to store committed unprocessed transactions. Both need to be on highly available storage with a preference for storage across Availability Zones (AZs), such as Elastic File System (Amazon EFS) or Windows File Server (FSx).
- The Amazon EC2 instance (or other Cloud instance) can be part of an Auto Scaling Group spread across AZs with minimum and maximum of one Amazon EC2 instance.
- Upon failure, the replacement Amazon EC2 instance of the replication product's administrator function is launched and communicates its IP address to the product's mainframe administrator function. The mainframe then starts communication with the replacement Amazon EC2 instance.
- Once the Amazon EC2 instance is restarted, it continues processing at the next logical restart point, using a combination of the LUW and Restart files.
- For production workloads, Treehouse Software recommends turning on Multi-AZ target and metadata databases.

Scalable Storage

- With scalable storage provided on most Cloud platforms, the customer pays only for what is used. The data replication product should require file-based storage for its files that can grow in size if target processing stops for an unexpected reason. For example, Amazon EFS, and Amazon FSx provide a serverless elastic file system that lets the customer share file data without provisioning or managing storage.

Analytics

- All top Cloud platform providers give customers the broadest and deepest portfolio of purpose-built analytics services optimized for all unique analytics use cases. Cloud analytics services allow customers to analyze data on demand, and helps streamline the business intelligence process of gathering, integrating, analyzing, and presenting insights to enhance business decision making.
- A data replication product should replicate data to several data sources that can easily be captured by various Cloud based analytics services. For example, mainframe database data can be replicated to the various Cloud 'buckets' in JSON, CSV, or AVRO format, which allows for consumption by the various Cloud analytic services. Bucket types include AWS S3, Azure BLOB Data, Azure Data Lake Storage, and GCP Cloud storage. Several other Cloud

analytics type services also support targets including Kafka, Elasticsearch, HADOOP, and AWS Kinesis.

- Kafka has become a common target and can serve as a central data repository. Most customers target Kafka using JSON formatted replicated mainframe data. Kafka can be installed on-premises, or using a managed Kafka service, such as the Confluent Cloud, AWS Managed Kafka, or the Azure Event Hub.

Monitoring

- Monitoring is a critical part of any data replication process. There are several levels of monitoring at various points in a data replication project. For example, each node of the replication including the mainframe, network communication, Cloud VM instances (such as EC2) and the target Cloud database service all can require a level of monitoring. The monitoring process will also be different in development or QA vs. a full production deployment.
- A data replication product should also have its own monitoring features. One important area to measure is performance and it is important to determine where any performance bottleneck is located. Sometimes it could be the mainframe process, the network, the transformation computation process, or the target database. A performance monitor helps to detect where the bottleneck is occurring and then the customer can drill down into specifics. For example, if the bottleneck is the input data, areas to examine are the mainframe replication product component performance, or the network connection. The next step is to monitor the area where the bottleneck is occurring using the data replication product's statistics, mainframe monitoring tools, or Cloud monitoring such as AWS CloudWatch.
- A data replication product should also allow the customer to monitor processing functions during the replication process. The data replication product should also have extensive logs and traces that allow for detailed monitoring of the data replication process and produce detailed replication statistics that include a numeric breakdown of processing statistics by table, type of operation (insert, update delete), and where these operations occurred (mainframe, or target database).
- CloudWatch collects monitoring and operational data in the form of logs, metrics, and events, providing customers with a unified view of AWS resources, applications, and services that run on AWS, and on-premises servers. You can use CloudWatch to set high resolution alarms, visualize logs and metrics side by side, take automated actions, troubleshoot issues, discover insights to optimize your applications, and ensure they are running smoothly.
- Some customers are satisfied with a basic monitoring that polls every five minutes, while others need more detailed monitoring and can choose polls that occur every minute.
- Amazon CloudWatch allows customers to record metrics for EC2 and other Amazon Cloud Services and display them in a graph on a monitoring dashboard. This provides visual notifications of what is going on, such as CPU per server, query time, number of transactions, and network usage.
- Given the dynamic nature of AWS resources, proactive measures including the dynamic re-sizing of infrastructure resources can be automatically initiated. Amazon CloudWatch alarms can be sent to the customer, such as a warning that CPU usage is too high, and as a result, an auto scale trigger can be set up to launch another EC2 instance to address the load. Additionally, customers

can set alarms to recover, reboot, or shut down EC2 instances if something out of the ordinary happens.

Disaster Recovery

- IT disasters such as data center failures, or cyber attacks can not only disrupt business, but also cause data loss, and impact revenue. Most Cloud platforms offer disaster recovery solutions that minimize downtime and data loss by providing extremely fast recovery of physical, virtual, and Cloud-based servers.
- A disaster recovery solution must continuously replicate machines (including operating system, system state configuration, databases, applications, and files) into a low-cost staging area in a target Cloud account and preferred region.
- Unlike snapshot-based solutions that update target locations at distinct, infrequent intervals, a Cloud based disaster recovery solution should provide continuous and asynchronous replication.
- Consult with your Cloud platform provider to make sure you are adhering to their respective best practices.
- Example: https://docs.aws.amazon.com/whitepapers/latest/disaster-recovery-workloads-on-aws/introduction.html).

Artificial Intelligence and Machine Learning

- Many organizations lack the internal resources to support AI and machine learning initiatives, but fortunately the leading Cloud platforms offer broad sets of machine learning services that put machine learning in the hands of every developer and data scientist. For example, AWS offers SageMaker, GCP has AI Platform, and Microsoft Azure provides Azure AI.
- Applications that are good candidates for AI or ML are those that need to determine and assign meaning to patterns (e.g., systems used in factories that govern product quality using image recognition and automation, or fraud detection programs in financial organizations that examine transaction data and patterns).

The list goes on…

- Treehouse Software and our Cloud platform and consulting partners can advise and assist customers in designing their roadmaps into the future, taking advantage of the most advanced technologies in the world.
- Successful customer goals are top priority for all of us, and we can continue to work with our customers on a consulting basis even after they are in production.

Your Mainframe Hybrid Cloud Partner

Treehouse Software is a global technology company and Technology Partner with AWS, Google Cloud, and Microsoft. The company assists organizations to accelerate digital transformation leveraging hybrid Cloud on the IBM Z platform.

Further reading: Treehouse Software's Mainframe-to-Cloud data replication tool is featured on the AWS Partner Network Blog: AWS Partner Network (APN) Blog: https://aws.amazon.com/blogs/apn/real-time-mainframe-data-replication-to-aws-with-tcvision-from-treehouse-software/.

For more information about Mainframe-to-Cloud data replication products and services from Treehouse Software, visit www.treehouse.com.

Treehouse SOFTWARE

2605 Nicholson Road, Suite 1230
Sewickley, PA 15143 | USA
1.724.759.7070
www.treehouse.com