

Using AWS Glue ETL jobs for Data migration from NoSQL DynamoDB to AWS Data Warehouse Redshift

**About Convosight**

**About Challenge**

Convosight is growing in business so it needs a robust, more secure and easily accessible and developer friendly system. So, Convosight is in place to bring all its existing external 3rd Party Services like Panoply to be migrated to cloud platforms like AWS(Amazon Web Services).

The data is kept in DynamoDB in the form of 52 tables. The data formats are String, Map, List and String Set etc. Currently Panoply is being used as SaaS for data warehousing. Panoply creates relational tables based on data types in DynamoDB. It also creates a sub table. A job runs after every 2 hours and the differential data from DynamoDB is inserted into relational mapping of tables.

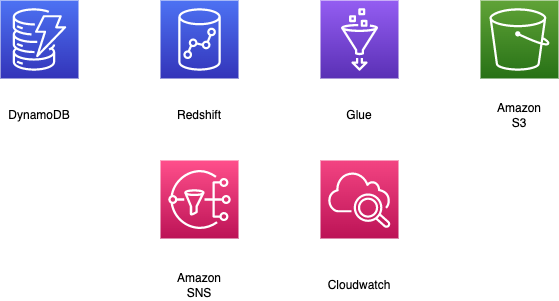
The major challenges we faced:

* support complex data types like list-of-strings (Creation of sub tables for these lists)
* support Schema changes (Addition of new columns at DynamoDB side)
* Migration of heavy table(few tables are 30GB-40GB)

Convosight is the world's first community management platform that uses the power of data analytics and machine learning to help admins and brands create, moderate, grow and leverage communities meaningfully.

Convosight is the world's first community management platform. 45,000 Facebook group admins trust Convosight to manage over 300M+ members, it helps them increase engagement, reduce SPAM and become community entrepreneurs.

**AWS Services used**



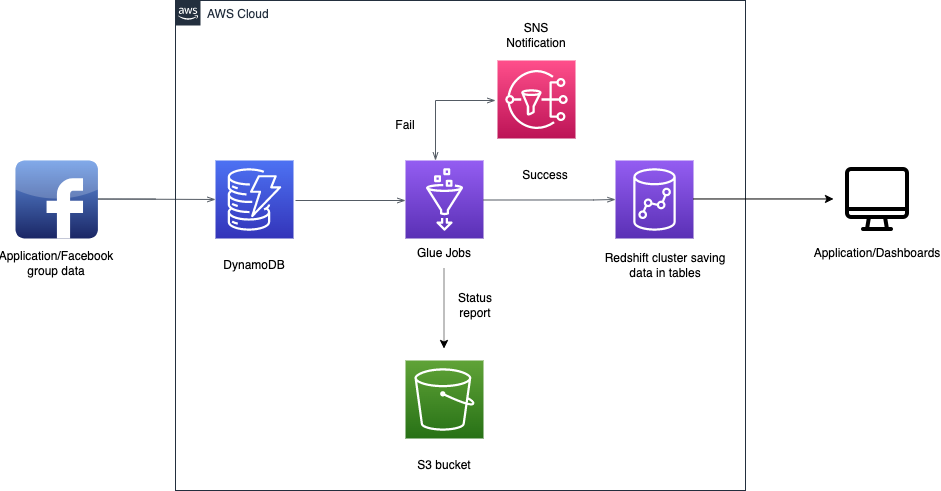
**Proposed Solution**

MIND team analyzed the problem and associated data and proposed ETL solution. Data is ingested from DynamoDB, transformed to suitable form so that it can be written to Redshift. This migration is a two-step process. The first step includes total migration of data up to date. The second step involves incremental data migration which runs four times a day on daily basis. The major transformation consists of casting to Redshift compatible datatype and breaking down complex DynamoDB datatype to simple scalar datatype for Redshift.

* Connection is established between DynamoDB and Redshift using Glue connection.
* Glue scripts takes up the DynamoDB data, converts it into dynamic data frame and does the following transformation.
* In the transformation function all string are trimmed to a length of 6000. All null datatype as casted to String datatype. All struct or map datatype are flattened.
* Any list datatype is unpacked and made into separate child table.
* The child table is referenced using key\_id column which is a hash created from DynamoDB partition key and sort key.
* CDC data is filtered on the basis of max created/updated time on last job run.
* Migrated all table along with sub tables for list datatype
* All table will support schema change in case of addition of new column to DynamoDB.
* All heavy table were migrated with CDC job set-up running twice a day.
* Remaining tables were set-up with CDC jobs running four times a day.

**Solution Outcome**

**Architecture Diagram**



**How AWS services helped in building ETL pipeline**

**AWS Glue for ETL processing**

AWS Glue is managed ETL service for data transformation and data ingestion and in this solution all the major transformations are done using Glue ETL scripts.

**Amazon Redshift for data warehousing**

Redshift is a fully managed, petabyte-scale data warehouse which is used to store the processed data from DynamoDB.

**Cloudwatch for monitoring**  
Amazon CloudWatch is a monitoring and observability service which keeps a tab on the set of services being used in the form of logs, metrics, and events.

**Amazon SNS for notifying failure or timeout of Glue job**

Amazon Simple Notification Service (Amazon SNS) is a managed service that provides message delivery from publishers to subscribers (also known as *producers* and *consumers*). Here, any job failure or timeout is intimated on mail using SNS**.**

**Amazon S3 for storage**

Amazon S3 is used to store the config file which will be used by the Glue jobs. The S3 Bucket is also used here to store the status report and temporary files created by the glue jobs, this helps in tracking the status of job completion and match the row counts.

**Amazon DynamoDB**

Amazon DynamoDB is fully managed NOSQL database which stores all the meta data of Facebook group data.

**About the Partner**

**MothersonSumi INfotech &Designs Ltd.**

MothersonSumi INfotech & Designs Limited (MIND), a SEI CMMI Level 5 IT services company and the IT back bone of Motherson group. MIND is a trusted technology partner to over 200 clients globally. Our value proposition is in our strength in specific Industry segments and years of experience in the areas of intelligent warehousing, Supply chain enablement, software application development, smart ERP customization, infra managed services, cloud, IoT & Analytics. MIND is serving customers in 41+ countries with a strong team of 1500+ professionals.