CDC changes from DynamoDB table to Redshift table:

The solutions below enable data replication between DynamoDB and Redshift in a generic way where schema changes are handled by Redshift process. A scheduled process keeps DynamoDB table and corresponding Redshift table in sync both in data and structure. Additional analytics processes can be built on these base tables as use case demands.  
  
At high level the steps are

* Create a Kinesis data stream (KDS) and turn on the data stream to KDS for tables to be replcated from DynamoDB.
* Create a materialized view in your Amazon Redshift cluster to consume live streaming data from the data stream.
* The streaming data gets ingested as a JSON payload. This JSON data is processed into Redshift tables via a Stored Procedure.

DynamoDB can stream multiple tables into a single stream. So, there will be one stream processed by Redshift to apply changes.  
  
  
Steps:

1. Create Kinesis Data Stream. (Example name: *redshift\_cdc*)
2. Create a Dynamo DB table and enable data streaming to Kinesis Data Stream created in step 1.
3. Create an IAM role with ability to read KDS streams and attach it to Redshift cluster. (Sample Role in appendix. Example name: *my\_streaming\_role*)
4. Create Redshift process.
   1. Create Stored Procedure on Redshift Cluster either in QEv2 or via any SQL tool. Create all procedure in attached files. (Procs\_DDB\_Redshift\_1.sql and Procs\_DDB\_Redshift\_2.sql)
   2. Verify below procedures are created
   3. ****
   4. Execute this procedure to create materialized view and schema replacing with IAM role, Account\_Number and KDS name. one time process.

**call public.sp\_ddb\_to\_redshift\_setup\_schema\_mv**(*'my\_streaming\_role','123456781234','redshift\_cdc')*;  
This will create necessary schema and materialized view to capture data from DynamoDB.   
  
d. Execute procedure below to create tables needed for replication process. one time process.  
**call public.sp\_ddb\_to\_redshift\_setup\_process\_tables();**

1. Procedure to replicate data. This is the scheduled process or can be run on demand:
   1. call public.sp\_ddb\_to\_redshift\_incremental\_refresh\_cdc();

Notes:

1. This is for CDC only for an existing table. In production scenario you may want to do a full copy/load via S3 and then use this process for ongoing changes.
2. For a new table this can capture data from initial changes if process is configured prior to new data.
3. You can run CDC with multiple tables pointing to same Kinesis Data Stream.
4. A new column is added to target table (dist\_Key) to track keys and distribution.
5. Target data table has all columns defined and stored as **varchar** to accommodate any future changes to attributes.