Datapipeline

Replicate Salesforce data into Redshift Using Datapipeline:

Performance tuning for Loading data into Redshift

DBSync Cloud Replication for Source is used for replication and synchronization of schema and data between Salesforce and Database namely Oracle, SQL Server, MySQL, Redshift and Cassandra.
We have often come across situation where in direct writes to redshift is quite slow and the requirement is to increase the performance of large data writes to redshift Data warehouse.

The goal of this white paper is to help you to understand how to increase the download performance into redshift by using an intermediate AWS RDS instance and push the whole load using AWS data pipeline.

We review in detail how to set up an AWS Data pipeline and steps to achieve clean download maintaining scalability and performance. After reading this whitepaper you will be able to make an educated decision and choose the solution that best fits your needs.

Setting up AWS Datapipeline for RDS MySQL to RedShift

- We shall make use of an intermediate RDS instance to load data into Redshift. To do so we need to first connect our Salesforce and any Staging Databases like RDS MySQL on AWS.
- Please find the below link on setting up Dbsync Cloud replication to connect to Salesforce and any AWS RDS instance like MySQL , Sql Server, Oracle etc.
  https://help.mydbsync.com/docs/pages/viewpage.action?pageId=53903406
- In your amazone console goto Analytics and select datapipeline. As per your requirement you can select the template from the "source" button in my case i need move data from my MySQL RDS to RedShift. Also based on your requirements you will be able to create deploy your custom templates for deploying datapipeline.
- As per your requirement you will be able to schedule through the console. I have scheduled for every week starting from 17-11-2015 also specified a convenient time so that we can make sure that there is no much load on the server.
- Set your database parameter and additional setting on the "parameter" section also you will be able to change the data type conversion etc. You will be allowed to modify your requirement in templates.
- Next step is "Pipeline Configuration" you can store you logs in s3 bucket also you have an option to disable to logs in it. Specify your IAM roles on the console.
- You can check and edit your configuration in "Edit in Architect"
- In Architect you have option to modify update and change your template properties. You have "Add activity" tab to add any type of activity you need to attach with pipeline configuration. Add Data Node for adding more source and targets into your pipeline. "Export" option used for exporting your template code
- Next step is you need to save the pipeline and Activate it.
- After activation you can monitor the logs while you click on the data-pipeline.