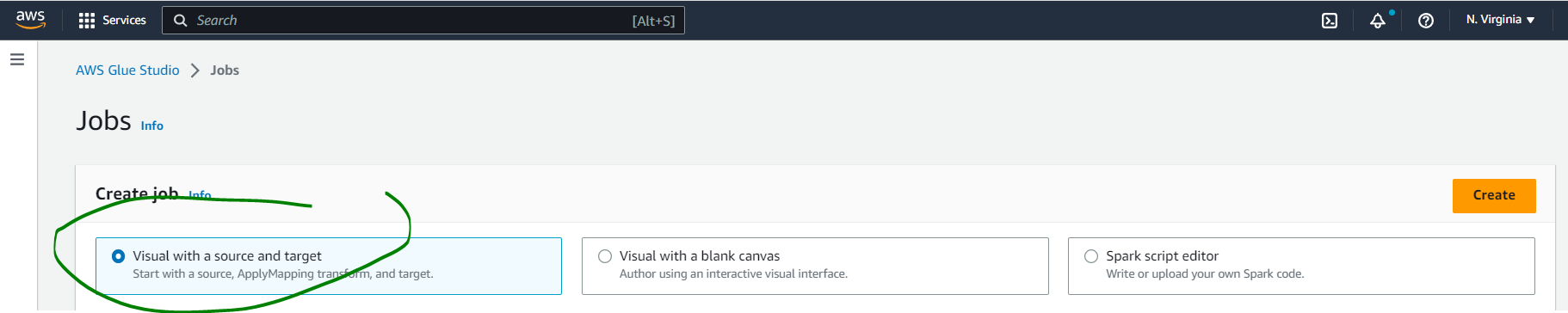
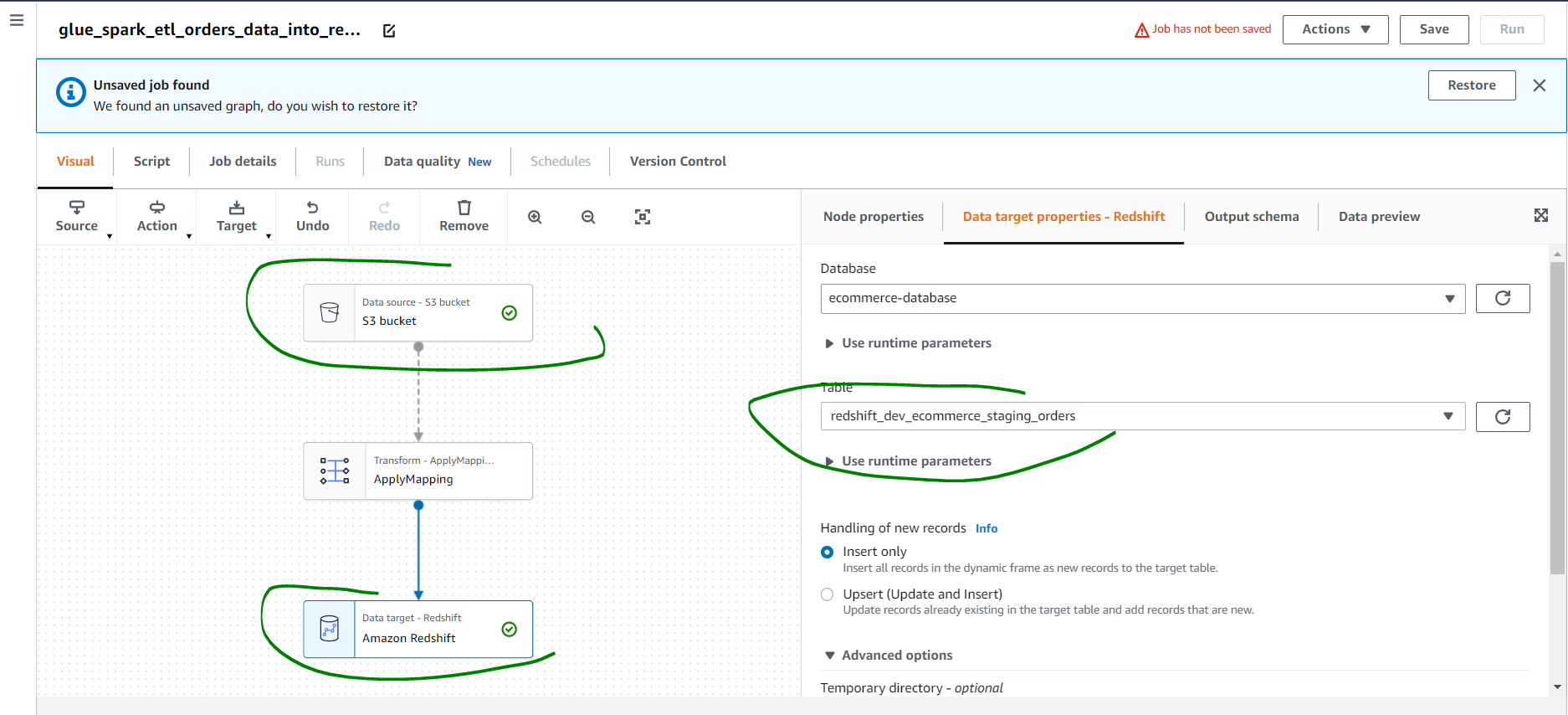
Since we plan to perform a simple / direct load from S3 staging area into Redshift staging area, we can use the Visual with a source and target option. A Spark ETL job is preferred over using a COPY command in a Python Shell job because we can enable Job Bookmark in the Spark ETL job and this is not possible in a Python Shell job. Enabling job bookmark will help us in not processing the same data from S3 staging area multiple times.

If we are using a Python Shell job for this, we might have to take care of Job bookmarking and we won’t be able to use the feature provided by AWS.





In the ApplyMapping section, we are not dropping any column.

Ensure you enable Job Bookmark in the Job Details section, customize Job Timeout, number of workers etc as needed. Associate an IAM Role that has required access on source S3 bucket and target Redshift table.

Run the Job. Verify data is loaded in the Redshift staging area. To verify Job bookmark is working, simply rerun the job and after success, you should not see an increase in the number of records in the Redshift staging table.