**Big Data Analytics: Amazon Reviews Report**

**Part 1: ETL of Amazon Review Datasets**

**Extract**

The data used for this analysis was collected from the Amazon customer review dataset website. I selected two of the datasets to look at in depth: video games and pet products.

**Transform**

The data was cleaned using pandas in Jupyter Notebook/Google Colab. A variety of dataframes were created following the required table schema provided. The appropriate columns were selected and the data types for each column were matched to the schema. Any calculations required for the schema were also applied.

**Load**

The tables for each of the datasets were loaded into AWS Relational Databases.

**Part 2: Are Vine Reviews Trustworthy?**

**ETL**

This analysis was performed on both datasets extracted in Part 1 of this report. A dataframe was created containing review id, star rating, number of helpful votes, total number of votes, and whether the reviewer was part of the Vine program or not.

First, I split the dataframe into two dataframes, one for vine reviewers and one for non-vine reviewers. I then compared the two dataframes across the following variables.

**Total Number of Reviews**

For both datasets there were far more reviews submitted by non-vine members (1,781,596 and 2,633,026) than vine members (4,290 and 10,215).

**Number of 5 Star Reviews**

For the video game dataset, 37% of vine reviews rated the product as 5 stars (1,607 out of 4290 reviews), while 58% of the non-vine reviews rated 5 stars (1,025,249 out of 1,781,596 reviews). For the pet products dataset, 42% of vine reviews rated the product as 5 stars (4343 out of 10,215 reviews), while 62% of the non-vine reviews rated 5 stars (1,640,940 out of 2,633,026 reviews). In both cases, a greater proportion of the non-vine reviews rated the product higher than vine reviews.

**Average Star Rating**

For the video game dataset, vine reviewers and non-vine reviewers gave almost identical average star ratings (4.07 and 4.06). This held true for the pet products dataset as well (4.07 and 4.14). So reviewers are rating the products fairly equally according to the averages.

**Number of Helpful Votes**

For the video game dataset, the vine reviews received 72% helpful votes (10,076 helpful votes and 14,064 total votes) while the non-vine reviews received 60% helpful votes (4,024,920 helpful votes and 6,696,252). For the pet products dataset, the vine reviews received 83% helpful votes (20,057 helpful votes and 23,943 total votes) while the non-vine reviews received 85% helpful votes (4,347,837 helpful votes and 5,099,623 total votes).

**Conclusion: Vine Reviews are Trustworthy (at least for these two datasets)**

Based on this analysis it appears that Vine reviews are trustworthy. While it appeared that non-vine members were providing more 5 star reviews than vine members, only looking at 5 star reviews is a very narrow focus and it is possible that vine members may not be rating products much lower than non-vine members. For example, vine members could be giving more 4 star reviews than non-vine members. This is supported by the fact that the average star ratings for vine and non-vine members were the same (around 4 stars). So that means that both groups are rating the products equally on average and supports the validity of Vine reviews.

The reviews of vine and non-vine members were also rated equally helpful for the pet products dataset, which also supports the trustworthiness of Vine reviews. However, there was a slight difference in helpful votes for the video game dataset with Vine reviews coming in slightly lower than non-vine reviews. This suggests that only looking at reviews in two product categories (video games and pet products) is not a large enough scope to truly determine if Vine reviews are trustworthy or not. Further analysis should be conducted on a larger sampling of the datasets to truly determine if Vine reviews are valid and consistent.