**DATA TRANSFORMATIONS**

**Dataset**: Video Game Sales With Ratings

**Date**: July 31, 2022

**Tools Used:**

* Azure HDInsight Cluster
* Azure file storage
  + Resource group: SNeudorf
  + Location: Canada Central
  + Storage account: sneudorfhdistorage
* Apache Hive 2.0

**Data Transformations Performed:**

Within Apache Hive 2.0:

1. Created table called, “sales\_genre”
2. Loaded the “sales” data into the “sales\_genre” table
3. Rounded the data in the “Global\_Sales” column to a whole number
4. Filtered the data to only show items in the “Critic\_Score” column that are greater-than zero
5. Ordered the data in the “Critic\_Score” column from highest to lowest

**Scripts Used:**

Below were the scripts used for the data transformations listed in the ‘Data Transformation Performed’ section above:

1.

CREATE TABLE sales\_genre

(

Genre string,

Global\_Sales float,

Critic\_Score float

);

2.

INSERT OVERWRITE TABLE sales\_genre

SELECT Genre,

Global\_Sales,

Critic\_Score

FROM sales;

3.

INSERT OVERWRITE TABLE sales\_genre

SELECT Genre,

ROUND(Global\_Sales) as Global\_Sales,

Critic\_Score

FROM sales\_genre;

4.

INSERT OVERWRITE TABLE sales\_genre

SELECT \*

FROM sales\_genre

WHERE Critic\_Score > 0;

5.

INSERT OVERWRITE TABLE sales\_genre

SELECT \*

FROM sales\_genre

ORDER BY Critic\_Score DESC;