Cleaning Data via Beeline in Hadoop:

Types of Analysis:

1. Time-Series Analysis
2. Natural Language Processing
3. Geospatial Analysis
4. Business Analysis
   1. Product Analysis, Sales, etc.

Using queries such as Most Common, Most popular, Most Reviewed, Most Time Played in a period of time

Data Types to Keep vs. Remove out of the 23 total cols.

Keep all Except:

Author.LastPlayed

Author.num\_reviews

Comment Count

Votes Funny

Votes Helpful

Weighted Vote Score(?)

--

What to do with the rest:

1. Language = Derived Geospatial Analysis of Assumed Review Location based on Region
2. Review text = NLP – What can we extract from the text? “Good” = good review, “Bad” = bad review.
3. Received for free = does this introduce bias? Cross check with Recommended. Are free games higher or lower rated than paid?
4. Recommended: Top 10 Recommended Titles, vs. Top 10 Most Reviewed Titles
5. Create a NEW data type called Genre for the Top 10 Recommended vs. Reviewed Titles?
6. Timestamp created – Is a unix timestamp… need to convert into our time lol wtf.. try almost 1mil for two records..