Data files:
File 1: reviews_Sports_and_Outdoors.json
File 2: meta_Sports_and_Outdoors.json

Create a dataframe with all the columns from “File_1” and add new columns, title and category from “File_1” based on same “asin” for the respective product.
In this dataframe, we will have the following columns:

reviewerID
asin
reviewerName
helpful
reviewText
overall
summary
unixReviewTime
reviewTime
title
price
brand
categories

Once we have this dataframe, we can select which column we want to work with. It seems to me that the price and brand could be good columns for recommending a relevant product!

Load the newly created dataframe in to Cassandra. It might be a good idea to get everything in a common workspace so that it can be accessible by all of us.

Preprocessing / feature engineering:
Group the data based on their name (title column) or asin and get the mean ratings.
Now, think about the situation, we may get lots of products with 5stars rating:

- What if only 1 or 2 persons have rated a product and gave a 5 stars rating to it!
- On the other hand, there could be a product that got a good number of ratings between 1 and 5 stars. The average in such situation will be less then 5 but the number of buyers would be higher!

This could be the case in our data, we need to check the products with number of ratings now. We can use groupby on title (or asin) and grab the ratings column. Instead of mean, we want count for this time. (now we have two new columns in our dataframe average_rat and n_rat)
Another thing, it would be a good idea to create another column with the length of the review. We may use this for ML part using NLP(if time permits)

At this stage, we need to explore our data and get some nice plots for the report.

Now our dataframe will look like, product as index along the rows and all other columns as predictors.

Continue….. will fill it up!