Greetings all,

I hope you are doing well. The purpose of this email is to report the project’s process and to clarify a few things about the data to make sure everything runs smoothly. Firstly, the “Date” features - “createdDate”, “lastLogin”, “dateScanned”, “finishedDate”, “modifyDate”, “pointsAwardedDate”, “purchaseDate” of “users” and “receipts” json files have additional property called “$date”. Although this is not a serious issue, it can mess up the data when we convert json files to relational tables for further query process if it is not handled properly. While I personally do not think the mentioned additional property is useful information and can erase it, I would like to check with you if it is important and what it means.

Then, I did exploratory analysis of the data to check the data quality such as duplicated data, incomplete fields, inconsistent formats, and different measurement units. The first thing to be discovered is that the date format is in unix milliseconds which is difficult to understand. Another thing is that there is a significant number of missing data points.

While there are several techniques to tackle the missing data problem, it will be a hard task when there are more than half of data missing. So, I would like to check if there is a way to collect these missing data again because sometimes, human error can lead to the problem of missing data. The missing data filling solutions should only be used at the last step.

Moreover, as I have mentioned above, the date time is unix-time milliseconds and it is necessary to convert it back to standard date time format. After changing it back, it has the information about hour, minute, and second. While this is not an entirely useless information, it is not very useful for querying. For example when we want to know the information about the most recent customer, we would go only as far as the date. It will only take extra memory to store this data. So I would like to know if I can remove such data to optimize the data assets.

Overall, the data has numerous unclear, unnecessary, and incomplete information. This can hugely affect the performance of our querying process. In addition the inclusion of unnecessary data as mentioned above can give us storage and scaling problems when we do production with a huge dataset. As a result, we need to eliminate all the unnecessary data and format our data to the simplest form from the data cleaning stage to have a smooth process at the later stages.

Best Regards

Htein Linn Thar