To: [michaelscott@fetchrewards.com](mailto:michaelscott@fetchrewards.com)

Subject: Update on the Business Questions –

Hi Mike,  
Hope you’re doing well, I had a bunch of quick questions that I hope you could give me an answer or at least point me to who I can get answers to these questions from.

Can you tell me a little bit about where you got this data from and what is the data generating process behind it? How this data is generated, by whom and how we store it. If you don’t have an answer please put me in touch with someone who knows this.

When I explored the data, I found a lot of issues like duplicate rows and lots of incomplete fields. I think this is something we need to work on to make a more streamlined process so we could probably automate this in the future going forward.

I think I can resolve some of these issues by

1. Knowing more about the data generating process
2. Understanding the attributes, we store and why they are important, also how they are related to each other
3. Knowing more about the tables in the source database where we picked this up from.

Some more information that would be helpful is knowing, who is going to use information? Is it the ML team or the Analytics team? Does the data need to be current or it’s okay if we store historical information?

I also have some concerns with performance and scaling going forward, If the data were to get too large we could possibly move to a distributed processing framework like Spark to get faster results.

If we want to automate this thing as much as possible and create an ETL pipeline of sorts, I would recommend that we work with the records as objects instead of strings and json because that would make consuming it, manipulation, and processing easier to read and understand for everybody. Since we control the entire pipeline right from generating the data, I think this should be relatively easier to do.

Please give me a call when you get the chance so I can go over these in detail.

Have a nice day!