

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

Subject: Update on data quality issue discovered during the data analysis task

Hi Courtney,

Hope you're having a great week!

I discovered data quality issues during my recent analysis of data. I had a few quick questions that I was hoping to find answers to. Can you please take a look and put me in touch with someone who can provide me with answers to these questions?

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 believe I can resolve some of these issues by

1. Understanding the data generation 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.
4. knowing who is going to use the information? Data science team or is it for reporting purposes? Does the data need to be current or do we need historical information?

From the solutions perspective, understanding the data generation process will help me examine the duplicate records which I found in the tables. This can be solved by introducing new business rules to check before inserting records in the table.

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 for quicker data processing and use cloud databases that offer auto-scaling. To automate the end-to-end data collection and storage process, we can create ETL pipelines as well.

Happy to schedule a call to go over these in detail.

Have a nice day!

Best,

Vishal