Fourth: Communicate with Stakeholders

For this question I shall write an email to a fake team lead/manager called James to inform about the data quality issues that were found.

Subject: Data Quality Issues with Receipts, Users and Brand Data

Body:

Hi James,

I hope all is well!

After conducting an extensive analysis for the records in Receipt, Users and Brands, I came across some discrepancies within the data. I discovered the data quality issues by ingesting the raw data into a warehouse and querying the data with automated and manual checks to see of any issues

1. Some question I have about the data are:
    a. When do we flag that a receipt needs a Fetch Review?
    b. What is the process of entering and updating brand information?
    c. Does barcode have a set standard format?
    d. How do we utilize user flagged items?
    e. What is the data retention policy?

2. To resolve the issue and to optimize the data set, I would like to inquire:
    a. Are Users, Receipts and Brands data being generated in real time or in batches?
    b. What are the acceptable values for certain fields in the brands, users and receipt databases?
    c. Which fields are required to be non-nullable?
    d. Are there any legacy systems that I would need to take into consideration?

3. Given the large volume of data there are some performance and scaling concerns that I would like to address. As the volume of the data increases, as it currently stands, there will be a lot of duplicate records and a large product catalog. To address this, I recommend implement a data deduplication process and sure that the indexing is optimized for scale. In addition, we would need to explore a more efficient way to handle user-flagged items, as well as invalid or inconsistent data like in barcode. These can be addressed by having some form of format validation and data consistency checks using triggers or constraints.

    Moreover, data caching via Redis can be used for frequently accessed data to reduce the load on the database and speed up response time. We would also set an appropriate cache expiration policy to ensure that cache does not become stale.

Furthermore, as the system scales, we can increase the number of servers to reduce bottleneck. We can implement a primary and secondary servers where primary servers handle the write transactions and use the secondary replicated servers for reading purposes to reduce the load on the write database. Also, we can implement load balance to allocate the load across multiple servers and horizontal scale the servers to handle increased traffic and data.

Lastly, if replicating servers are deemed too costly, I would recommend setting up an ETL pipelines to periodically move data from the operational database to a data warehouse that is designed for more OLAP purposes. This is will reduce the performance impact of the transactional databases as we scale.

I will work with the team to define the clear data entry and validation guides to prevent these issues from reoccurring, as well as discuss scaling for performance enhancements. Please let me know if you have any further questions.

Thanks.

Kind Regards,
Affan Rashdi