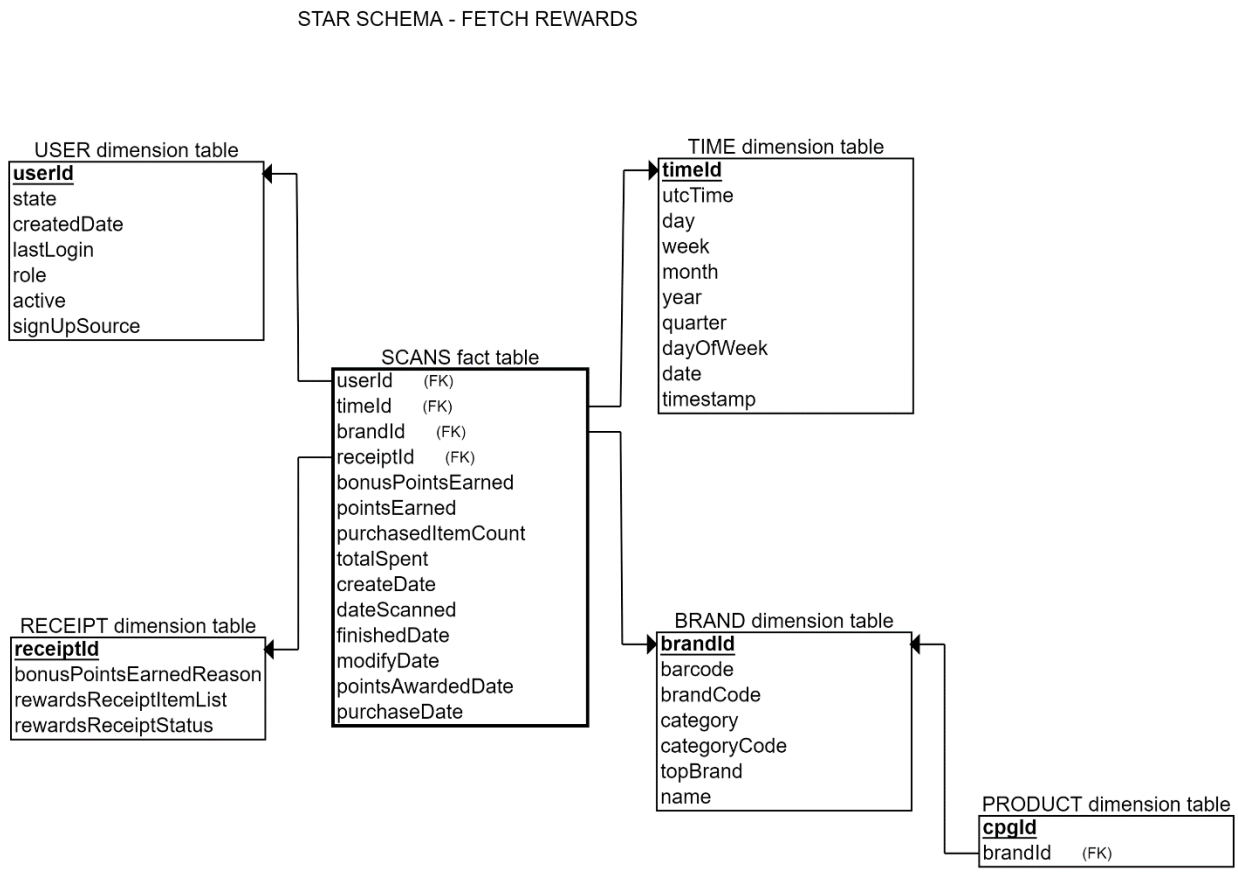
Review Existing Unstructured Data and Diagram a New Structured Relational Data Model

Following Kimball's methodology, the unstructured JSON file contents were organized into a simple Snowflake-schema structure with 5 Dimension Tables and one Fact Table. Each of these Dimension Tables was then assigned a surrogate key to uniquely identify rows of their data along with ensuring the contents of JSON were divided based on if the attributes were a dimension or a measure. The measures were then added to the Fact Table 'SCANS' that stores all surrogate keys of its dimensions along with additive and semi-additive facts.

Additionally, the Time Dimension Table was populated with more details to achieve granularity, i.e., to answer questions on different levels of grain, e.g., daily, monthly, quarterly, annual values for sales and transactions.



ERDPlus. (2021). Retrieved 12 August 2021, from https://erdplus.com/

Write a query that directly answers a predetermined question from a business stakeholder

NOTE:

1. Tables are temporarily named with the prefix FR.
2. Date inserted into the *FR_time* table was cleaned during ETL phase and converted from UNIX milliseconds to its relevant columns

What are the top 5 brands by receipts scanned for most recent month?

```
WITH
    earliestmonth
AS
(
    SELECT
        strftime('%m',date)
    FROM
        FR_time
    ORDER BY date DESC
    LIMIT 1
)

SELECT
    b.name AS "Top 3 Brand Names"
FROM
    (
    SELECT
        brandId
    FROM
        (
        SELECT
            brandId, COUNT(brandId) AS pop
        FROM
            FR_scans sc
        INNER JOIN
            (
            SELECT
                timeId
            FROM
                FR_time
            WHERE
                strftime('%m',date) IN
                (
                SELECT
                    *
                FROM
                    earliestmonth
                )
            ) sub
        ON sc.scannedDate = sub.timeId
        GROUP BY brandId
        )
    ORDER BY pop DESC
    LIMIT 5
    ) id
INNER JOIN
    FR_brand b
ON id.brandId = b.brandId
;
```

Evaluate Data Quality Issues in the Data Provided

https://github.com/saraiyash/FR-Data-Analyst/blob/main/fetch-rewards-quality-checks.ipynb

Conclusion:

Criteria for measuring data quality:

1. Accuracy

a. Duplication of records was noticed in the dataset. Total duplicate records in the dataset = 283. These records were dropped eventually. However, this could be an indicator of stale or inaccurate data.

ACTION - Reach out to the business stakeholders as well as the source system teams about the duplicates.

b. Almost all the other columns mentioned in the data dictionary met the required criteria except for the role column. The role column is a constant set to "consumer"; however, we can also see that some rows have "fetch-staff" value present.

ACTION - Ask the business stakeholders whether this is expected. Presumably, some of the actual FR staff have accessed this data and hence, this value is assigned to the records edited by the FR staff.

2. Relevancy

a. All the columns are relevant with respect to the user/consumer. No issues found.

3. Completeness

a. Incomplete records were found as there were multiple columns with NA/NaN and NULL values present. This could be a problem from the source of the dataset itself.

ACTION - Ask the source system team to investigate this issue with missing values or data.

4. Timeliness

a. Based on the 2 columns that provide date-based information, the latest update in the dataset is from 2021-03-05.

ACTION - Need confirmation on whether this is the latest cut-off date for the dataset or if more recent data somehow went missing, pending further investigation.

5. Consistency

a. No major issues with consistency of the data. Minor issue - date attributes are present in UNIX time. Need to confirm whether this is expected and consistent with other datasets. All the other files also were in the same UNIX time format which makes this a non-issue.

ACTION - Ask the business stakeholders and the source system team why the date format is set to UNIX time format.

Communicate with Stakeholders

Hi team,

Hope you are doing well!

Based on my preliminary analysis of the dataset, I have a few observations that I would like confirmed from both you as the business stakeholders and the source system teams.

Observations:

1. Duplicate records: There were around 283 duplicate records in the dataset of about 495 records. These records were eliminated during the exploratory analysis. However, the presence of these records may mean that the dataset was stale or inaccurate.

2. Role as "fetch-staff": A few records in the *role* column were set to *fetch-staff*. Although this seems like a non-issue as this could easily be testing data, I wanted to confirm once more as the data dictionary only mentions one constant value of *consumer*. Additionally, if this is not an issue, we could also add a small note about the presence of *fetch-staff* value in the data dictionary.

3. NA, NaN, Null values: About 51 fields in the entire dataset were populated with non-existent values. This could be an indicator of missing or incomplete dataset. Once again, I need to confirm whether these missing values are expected or not.

4. Date columns in UNIX Time format: Lastly, all the date columns were consistently set to UNIX Time format which matches the consistency quality check. However, I wanted to confirm whether this is the best strategy for storing date values in this format. If the format is changed to ISO 8601 [YYYY-MM-DD], this would greatly improve readability.

5. Cut-off date: Based on the latest date values in the dataset, the cut-off date for it seems to be 2021-03-05. The cut-off date being larger than this date could indicate data loss.

Moving on, I have additional queries with respect to the implementation of the data warehouse.

1. I wanted to know the size of the data that the DW would have to handle, since a traditional DW can get inflexible, and scalability could become an issue. To tackle this, we could consider a cloud DW which would tackle these issues, however, that could result in additional concerns surrounding Cloud Security and Data Governance.

2. In terms of a traditional DW, to avoid a single-point-of-failure, we could deploy a synchronous configuration. This would mean eradication of stale data and ensuring data is always available to the relevant stakeholders leading to less down-time.

3. In terms of possessing complementary tools, I was curious about what BI tools and ETL engines are in place already. Having an easy-to-use drag-and-drop interface to build data warehouse pipelines would cut down the resource costs in terms of both configuration and maintenance.

I look forward to your response. Have a great day!

Regards,

Yash