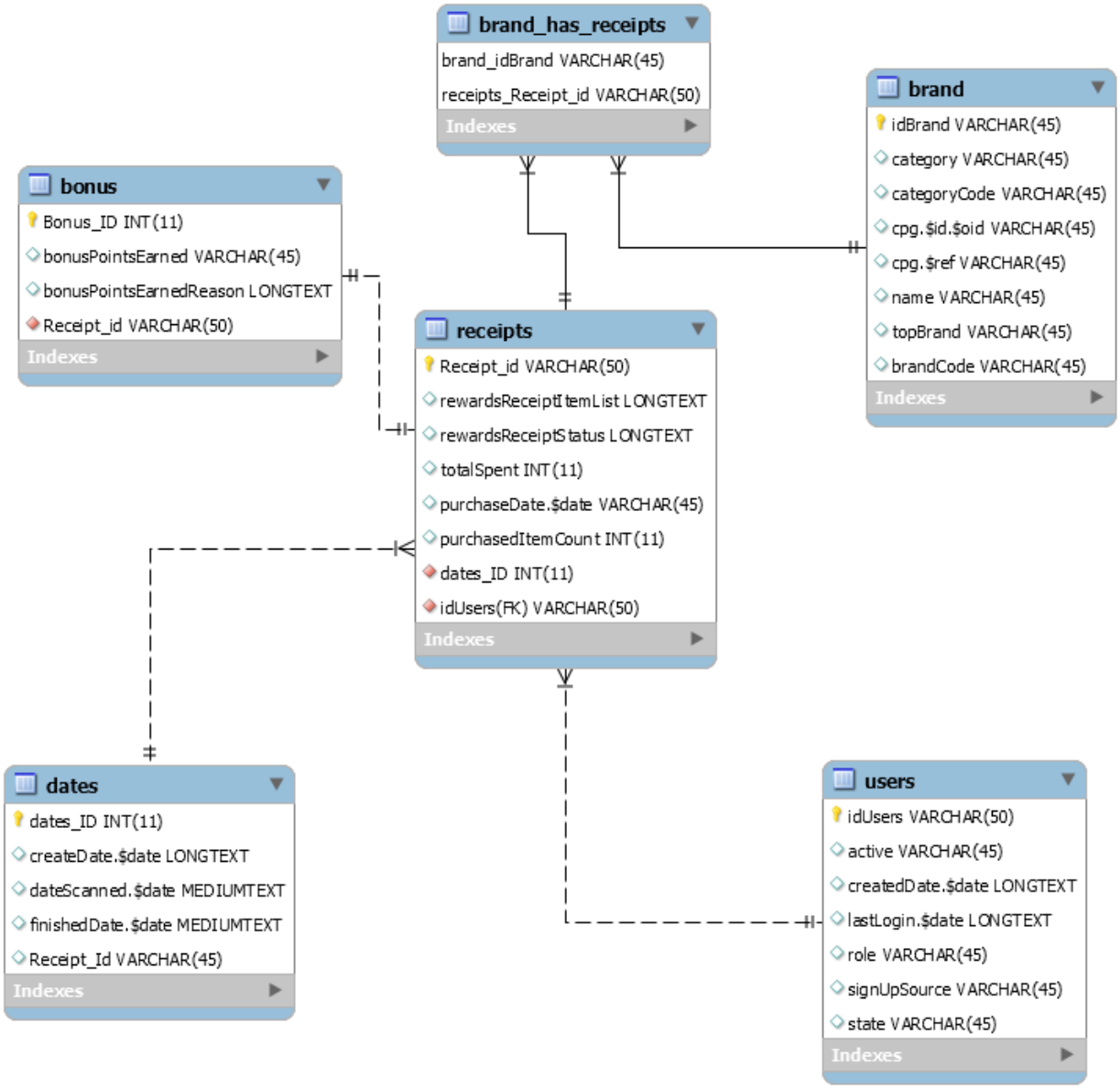
Fetch Rewards Analytics Engineer Assessment

Before moving ahead, I have converted all JSON files to CSV for convenience and to import in MySQL workbench

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



I have created ER model with the help of MySQL workbench using ETL queries and then reverse engineering it in the platform.

To make it simpler , I have normalized the data and divided receipts CSV into three tables I.e receipt, date and bonus.

Relationships:

Each user will have multiple receipts

Each date will have multiple receipts

Each receipt is unique

Brand will have multiple receipts, or many receipts will have multiple brands

Each receipt will have one bonus(considering you get only one bonus point after uploading the receipt)

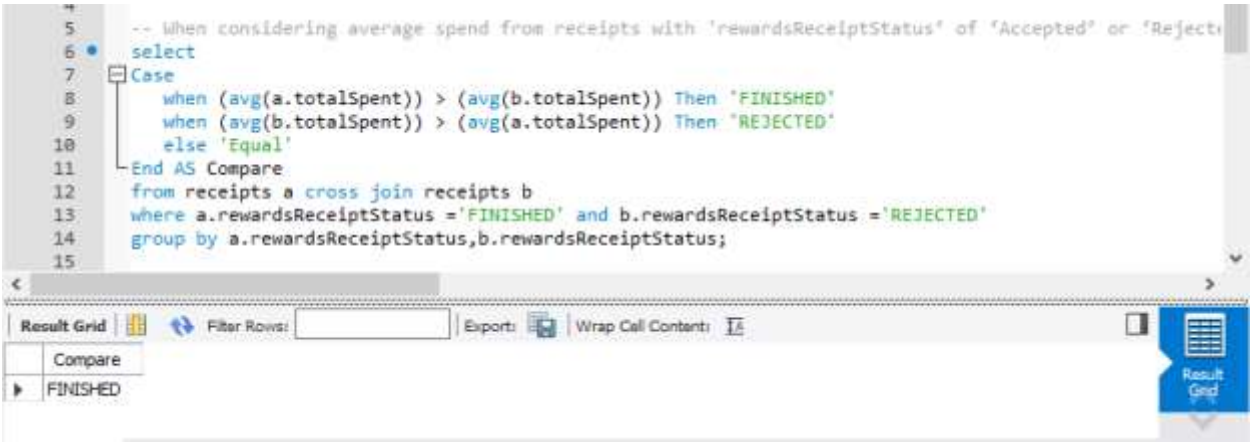
Second: Write a query that directly answers a predetermined question from a business stakeholder

Write a SQL query against your new structured relational data model that answers one of the following bullet points below of your choosing.

Even though one question was expected I tried to answer all of the questions mentioned

1) When considering average spend from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?

In this query I got Finished as greater but if you can run it in the systems you will get answer according to the data that loads.



2) What are the top 5 brands by receipts scanned for most recent month?
I have included auto increment date id in both table… you can use the same query without join if you do not decide to separate Receipts in three parts

```
-- What are the top 5 brands by receipts scanned for most recent month?
select brand.name as Brand, Count(receipts.Receipt_id) as Total_Receipt from brand
left join receipts on brand.Receipt_id = receipts.Receipt_id
inner join dates on receipts.dates_ID = dates.dates_ID -- asssuming dates and receipts have common column
group by brand.name
having datepart(mm,date.dateScanned.$date) =month(getdate())
order by Total_Receipt desc limit 5;

-- How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the

select*, LAG(Count(receipts.Receipt_id)) OVER (PARTITION BY brand.name ORDER BY month(getdate())
as previousmonth
FROM brand
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

idBrand	category	categoryCode	cpg.$id.$oid	cpg.$ref	name	topBrand	brandCode	idUsers
54943462e4b07e684157a532	true	1418998882381	1614963143204	fetch-staff	****	****	NULL	NULL
55308179e4b0eabd8f99caa2	true	1429242233186	1525713820003	consumer	****	WI	NULL	NULL
[illegible]	true	[illegible]	[illegible]	[illegible]	****	[illegible]	NULL	NULL

3) Which brand has the most spend among users who were created within the past 6 months?

I have included auto increment date id in both table… you can use the same query without join if you do not decide to separate Receipts in three parts

I have used curdate() function to get the current date

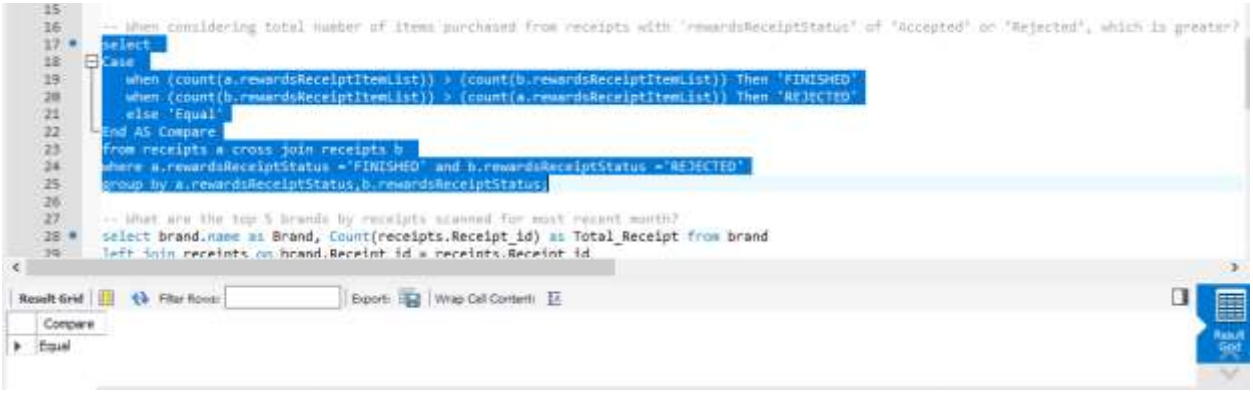
```
-- Which brand has the most spend among users who were created within the past 6 months?
select brand.idBrand as BrandID, brand.name as Brand, sum(receipts.totalSpent) as TotalSpent, date.createDate.$date as Date from brand
inner join receipts on brand.Receipt_id = receipts.Receipt_id
inner join dates on receipts.dates_ID = dates.dates_ID -- asssuming dates and receipts have common column
where Date > curdate() - interval (dayofmonth(curdate()) - 1) day - interval 6 month
group by BrandID, Brand
order by TotalSpent desc limit 1;
```

4) Which brand has the most transactions among users who were created within the past 6 months?

```
-- Which brand has the most transactions among users who were created within the past 6 months?
select brand.idBrand as BrandID, brand.name as Brand, count(date.dateScanned.$date) as No_of_Transaction from brand
inner join receipts on brand.Receipt_id = receipts.Receipt_id
inner join dates on receipts.dates_ID = dates.dates_ID -- asssuming dates and receipts have common column
where date.dateScanned.$date > curdate() - interval (dayofmonth(curdate()) - 1) day - interval 6 month
group by BrandID, Brand
order by No_of_Transaction desc limit 1;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

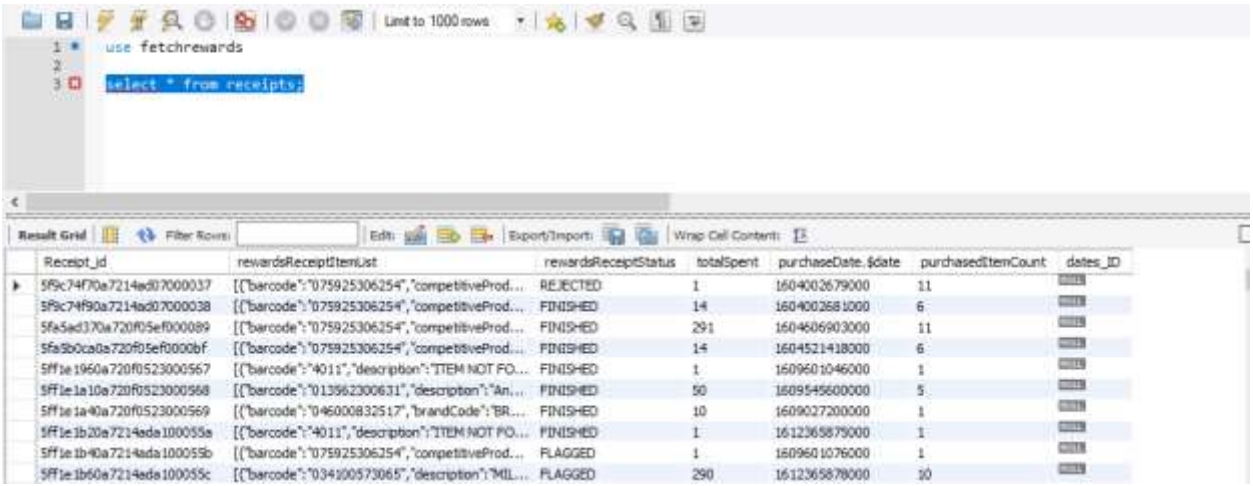
5) When considering total number of items purchased from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?
I have included count because Receipt item list does not contain numbers they have list of items. If you have number, you may use sum.



Third: Evaluate Data Quality Issues in the Data Provided:

I have decided to quick quality check using SQL. One can do Profiling with the help of Python to get more in-depth idea.

1) All ids such as Brand id, Receipt id is not in one format. Because of this reason I have decided to include to include VARCHAR() while loading the data.
2) All Dates are not in one format. We cannot convert it into DDMMYYYY or any other format



3) There are many “” values instead of NULL values this indicate it does not contain anything. We can replace these values with “NULL” in python.

```
1  use fetchrewards
2
3  select * from brand;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

idBrand	category	categoryCode	cpg.$id.$oid	cpg.$ref	name	topBrand	brandCode
54943462e4b07e684157a532	true	1418998882381	1614963143204	fetch-staff	****	****	NULL
55308179e4b0eabd8f99caa2	true	1429242233186	1525713820003	consumer	****	WI	NULL
5964eb07e4b03efd0c0f267b	true	1499785991771	1614884869770	fetch-staff	****	IL	NULL
59c124bae4b0299e55b0f330	true	1505830074302	1612802578117	fetch-staff	****	WI	NULL
5a43c08fe4b014fd6b6a0612	true	1514389647059	1613146957155	consumer	****	****	NULL
5e27526d0bdb6a138c32b556	true	1579635309795	****	consumer	Google	WI	NULL
5f2068904928021530f8fc34	true	1595959440905	1612452605375	fetch-staff	Email	WI	NULL
5fa32b4d898c7a11a6bcebce	true	1604528973309	1614842518047	fetch-staff	Google	AL	NULL
5fa41775898c7a11a6bcef3e	true	1604589429396	1614873722026	fetch-staff	Email	****	NULL
5fb0a078be5fc9775c1f3945	true	1605410936818	****	consumer	Google	AL	NULL

Fourth: Communicate with Stakeholders

Dear XYZ

Hope you are doing well!

I am writing this email regarding the recent data task. I have created data model and used ETL queries to answer business questions mentioned in the meeting. During the task, I observed a few irregularities in the data and would like to put light on them. They are as follows:

1) While importing the data in SQL, I observed that there are many irregularities in data. I wanted to make sure the source of the data. We have used JSON files to load but are there any other sources apart from that? We would like to check authenticity of date format and null values with different sources.
2) I would like to know if there are more CSVs apart from these three. It would be great to get more information to solve the business problems.
3) How would you like to view the data? Do they want to see it in a tabular format, interactive graphs and charts, and trending tables?
4) Are there any new KPIs dimension to view more detailed information that we can focus?

We can jump on the call to discuss this further. I would like to schedule 1 hr with the team. Please let me know your availability so we can connect!

Thanking you in advance!

Regards,

Chinmay Arolkar

Please refer to SQL files for the code.