MIS 443 Group Project 1: SQL Challenge

Project: Data Bank Case Study Analysis

The objective of this project is to apply SQL skills to analyze real-world datasets by addressing practical business questions from the Data Bank case study. The project involves writing SQL queries, extracting meaningful insights, and presenting findings.

Case Study Overview: Data Bank

Data Bank is a digital-only banking platform that integrates banking activities and cloud data storage allocation. Customers' storage limits are linked to their account balances.

Data Sources

1. Regions Table - Contains information about banking nodes worldwide.
2. Customer Nodes Table - Tracks customer allocations to banking nodes.
3. Customer Transactions Table - Stores deposits, withdrawals, and purchases.

Deliverables

A. Customer Nodes Exploration

How many unique nodes are there on the Data Bank system?

SQL Query:

```
6 ⌄  select
7         count(distinct node_id) as unique_nodes
8         from data_bank.customer_nodes
9     ;
```

unique_nodes 🔒 bigint
5

Insight: The system operates with 5 distinct banking nodes globally, likely corresponding to regional divisions for data security and customer distribution. This is crucial as it ensures financial data protection and balanced customer distribution across regions.

What is the number of nodes per region?

SQL Query:

```
13 ∨  select regs.region_name,
14     count(distinct cns.node_id) as number_of_nodes
15     from data_bank.regions regs
16     inner join data_bank.customer_nodes cns
17     on regs.region_id=cns.region_id
18     group by regs.region_name
19     ;
```

	region_name character varying (9) 🔒	number_of_nodes bigint 🔒
1	Africa	5
2	America	5
3	Asia	5
4	Australia	5
5	Europe	5

Insight: This suggests a balanced infrastructure, ensuring redundancy and security across all regions. If the regions are misnamed, it could lead to issues in tracking financial activities properly, which should be corrected for accurate data analysis.

How many customers are allocated to each region?

SQL Query:

```
23 ∨  select regs.region_name,
24     count(cns.customer_id) as customer_each_region
25     from data_bank.regions regs
26     inner join data_bank.customer_nodes cns
27     on regs.region_id=cns.region_id
28     group by regs.region_name
29     ;
```

	region_name character varying (9) 🔒	customer_each_region bigint 🔒
1	America	735
2	Australia	770
3	Africa	714
4	Asia	665
5	Europe	616

Insight: The customer distribution across regions is balanced, with Australia (Oceania) having the highest count at 770, while Europe has the lowest at 616. This suggests that while overall distribution is even, there may be stronger engagement in Australia, and an opportunity exists to implement targeted marketing strategies in Europe to boost customer acquisition and retention.

How many days on average are customers reallocated to a different node?

SQL Query:

```sql
33  with node_days as(
34    select
35    customer_id,
36    node_id,
37    end_date - start_date as days_in_node
38    from data_bank.customer_nodes
39    where end_date != '9999-12-31'
40    group by customer_id, node_id, start_date, end_date
41    )
42    , total_node_days as (
43    select
44    customer_id,
45    node_id,
46    sum(days_in_node) as total_days_in_node
47    from node_days
48    group by customer_id, node_id
49    )
50
51    select round(avg(total_days_in_node)) as avg_node_reallocation_days
52    from total_node_days
53    ;
```

avg_node_reallocation_days 🔒 numeric
1	24

Insight: An average of 24 days for customer reallocation to a different node suggests that customers are moved regularly, likely for security and system optimization. This frequency helps protect customer data from being stored in one location for too long but might cause minor disruptions for customers if they rely on a stable node for their banking activities.

What is the median, 80th, and 95th percentile for reallocation days per region?

SQL Query:

```sql
with node_days as (
select
cn.customer_id,
cn.region_id,
cn.node_id,
cn.end_date - cn.start_date as days_in_node
from data_bank.customer_nodes cn
where cn.end_date != '9999-12-31'
)
select
r.region_name,
nd.region_id,
percentile_cont(0.5) within group (order by nd.days_in_node) as median_days,
percentile_cont(0.8) within group (order by nd.days_in_node) as p80_days,
percentile_cont(0.95) within group (order by nd.days_in_node) as p95_days
from node_days nd
inner join data_bank.regions r
on nd.region_id = r.region_id
group by r.region_name, nd.region_id
order by nd.region_id
;
```

	region_name character varying (9)	region_id integer	median_days double precision	p80_days double precision	p95_days double precision
1	Australia	1	15	23	28
2	America	2	15	23	28
3	Africa	3	15	24	28
4	Asia	4	15	23	28
5	Europe	5	15	24	28

Insight An average of 24 days for customer reallocation indicates that customers are moved to a new node regularly, likely for security and system optimization. This ensures that data is not stored in one location for too long, but it may cause minor disruptions for customers who prefer stability in their banking services.

B. Customer Transactions Exploration

What is the unique count and total amount for each transaction type?

SQL Query:

```
81  select
82    txn_type,
83    count(*) as unique_count,
84    sum(txn_amount) as total_amount
85    from data_bank.customer_transactions
86    group by txn_type
87    order by txn_type
88    ;
```

	txn_type character varying (10)	unique_count bigint	total_amount bigint
1	deposit	2671	1359168
2	purchase	1617	806537
3	withdrawal	1580	793003

Insight: Deposits are the most common transaction type, making up the largest share of total transaction volume. This suggests that customers primarily use Data Bank as a savings or funding account. Since purchases and withdrawals are relatively lower, customers may be holding their funds in the system rather than spending frequently. This trend could indicate opportunities to encourage more spending activity, such as cashback promotions or rewards for purchases.

What is the average historical deposit count and amount for all customers?

SQL Query:

```
92 ⌄  select
93     round(avg(txn_count)) as avg_deposit_count,
94     round(avg(txn_amount)) as avg_deposit_amount
95     from (
96     select
97     customer_id,
98     count(customer_id) as txn_count,
99     avg(txn_amount) as txn_amount
100    from data_bank.customer_transactions
101    where txn_type='deposit'
102    group by customer_id
103    ) as deposit
104    ;
```

	avg_deposit_count numeric	avg_deposit_amount numeric
1	5	509

Insight: On average, customers deposit 5 times with an average deposit amount of $509. This indicates that customers deposit money regularly, likely in alignment with payroll cycles (monthly or bi-weekly). If Data Bank wants to increase deposit frequency or amounts, they could consider offering higher interest rates for larger balances, automatic deposit incentives, or tiered savings plans.

How many customers make more than one deposit and at least one purchase/withdrawal per month?

SQL Query:

```sql
110 v with monthly_transactions as (
111   select
112   customer_id,
113   date_part('month', txn_date) as mth,
114   sum(case when txn_type='deposit' then 1 else 0 end) as deposit_count,
115   sum(case when txn_type='purchase' then 1 else 0 end) as purchase_count,
116   sum(case when txn_type='withdrawal' then 1 else 0 end) as withdrawal_count
117   from data_bank.customer_transactions
118   group by customer_id, date_part('month', txn_date)
119   )
120
121   select
122   mth,
123   count(distinct customer_id) as customer_count
124   from monthly_transactions
125   where deposit_count>1
126   and (purchase_count >= 1 or withdrawal_count >= 1)
127   group by mth
128   order by mth
129   ;
```

	mth double precision 🔒	customer_count bigint 🔒
1	1	168
2	2	181
3	3	192
4	4	70

Insight: The number of customers making multiple deposits and transactions peaked in the first three months but declined in month 4. This could indicate seasonal trends, economic factors, or changes in promotional incentives.

Final Summary

The analysis reveals customer behavior trends, security policies, and banking behaviors, revealing areas for improvement. It also presents growth strategies through customer engagement, marketing, and financial forecasting. The report emphasizes the business impact of SQL-based data analysis and suggests Data Bank should focus on understanding regional policies, addressing security concerns, and catering to customer preferences for effective improvements.