

Introduction to the case study Data Bank

There is a new innovation in the financial industry called Neo-Banks: new aged digital only banks without physical branches.

Danny thought that there should be some sort of intersection between these new age banks, cryptocurrency and the data world…so he decides to launch a new initiative - Data Bank!

Data Bank runs just like any other digital bank - but it isn’t only for banking activities, they also have the world’s most secure distributed data storage platform!

Customers are allocated cloud data storage limits which are directly linked to how much money they have in their accounts. There are a few interesting caveats that go with this business model, and this is where the Data Bank team need help!

The management team at Data Bank want to increase their total customer base - but also need some help tracking just how much data storage their customers will need.

This case study is all about calculating metrics, growth and helping the business analyse their data in a smart way to better forecast and plan for their future developments!

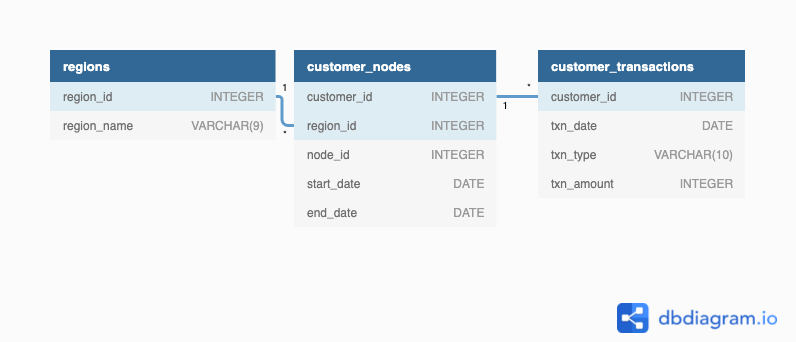
**Key Concepts:**

* **Neo-Banks**: Digital-only banks that operate without physical branches.
* **Distributed Data Storage Platform**: World-class secure storage system linked to customer balances.
* **Customer Data Allocation**: Customers receive data storage based on the amount in their accounts, blending banking with data management.

Available Data

The Data Bank team have prepared a data model for this case study as well as a few example rows from the complete dataset below to get you familiar with their tables.

**Entity Relationship Diagram**



*Before we dive into the analysis, let’s take a moment to understand the different tables that were eventually created in the database.*

**Table 1: Regions**

Just like popular cryptocurrency platforms - Data Bank is also run off a network of nodes where both money and data is stored across the globe. In a traditional banking sense - you can think of these nodes as bank branches or stores that exist around the world.

This regions table contains the region\_id and their respective region\_name values.

| **region\_id** | **region\_name** |
| --- | --- |
| 1 | Africa |
| 2 | America |
| 3 | Asia |
| 4 | Europe |
| 5 | Oceania |

**Table 2: Customer Nodes**

Customers are randomly distributed across the nodes according to their region - this also specifies exactly which node contains both their cash and data.

This random distribution changes frequently to reduce the risk of hackers getting into Data Bank’s system and stealing customer’s money and data!

**Below is a sample of the top 10 rows of the data\_bank.customer\_nodes**

| **customer\_id** | **region\_id** | **node\_id** | **start\_date** | **end\_date** |
| --- | --- | --- | --- | --- |
| 1 | 3 | 4 | 2020-01-02 | 2020-01-03 |
| 2 | 3 | 5 | 2020-01-03 | 2020-01-17 |
| 3 | 5 | 4 | 2020-01-27 | 2020-02-18 |
| 4 | 5 | 4 | 2020-01-07 | 2020-01-19 |
| 5 | 3 | 3 | 2020-01-15 | 2020-01-23 |
| 6 | 1 | 1 | 2020-01-11 | 2020-02-06 |
| 7 | 2 | 5 | 2020-01-20 | 2020-02-04 |
| 8 | 1 | 2 | 2020-01-15 | 2020-01-28 |
| 9 | 4 | 5 | 2020-01-21 | 2020-01-25 |
| 10 | 3 | 4 | 2020-01-13 | 2020-01-14 |

**Table 3: Customer Transactions**

This table stores all customer deposits, withdrawals and purchases made using their Data Bank debit card.

| **customer\_id** | **txn\_date** | **txn\_type** | **txn\_amount** |
| --- | --- | --- | --- |
| 429 | 2020-01-21 | deposit | 82 |
| 155 | 2020-01-10 | deposit | 712 |
| 398 | 2020-01-01 | deposit | 196 |
| 255 | 2020-01-14 | deposit | 563 |
| 185 | 2020-01-29 | deposit | 626 |
| 309 | 2020-01-13 | deposit | 995 |
| 312 | 2020-01-20 | deposit | 485 |
| 376 | 2020-01-03 | deposit | 706 |
| 188 | 2020-01-13 | deposit | 601 |
| 138 | 2020-01-11 | deposit | 520 |

Case Study Questions

The following case study questions include some general data exploration analysis for the nodes and transactions before diving right into the core business questions and finishes with a challenging final request!

The case study questions are grouped into:

**A. Customer Nodes Exploration**

**B. Customer Transactions**

**C. Data Allocation Challenge**

**E. Extra Challenge**

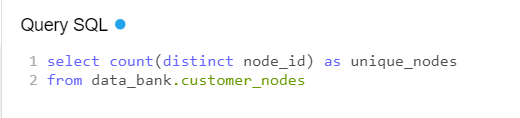
**F. Extension Request**

**Goals:**

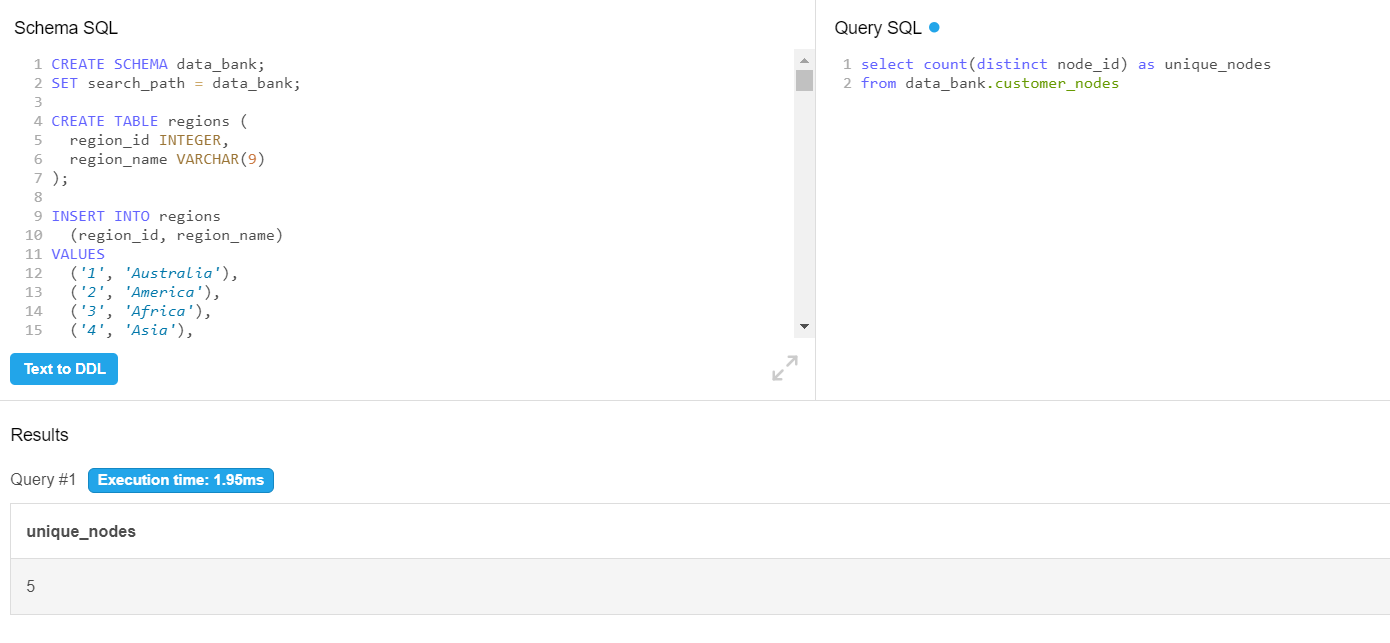
* **Customer Growth**: The Data Bank team is focused on increasing their customer base.
* **Data Storage Management**: The team needs help tracking the **data storage** needed for customers based on their account balances.

A. Customer Nodes Exploration

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

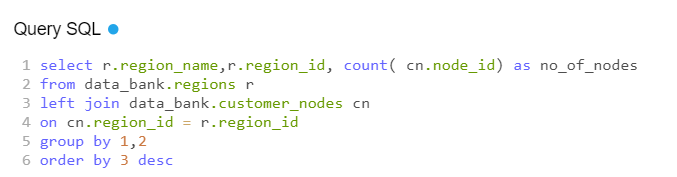
****

**Output-**



This query counts the number of distinct nodes (node\_id) in the **customer\_nodes** table, providing the total number of unique nodes within the Data Bank system i.e 5.

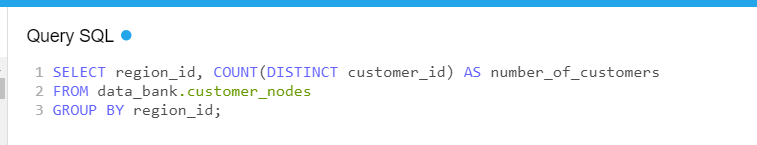
1. **What is the number of nodes per region?**

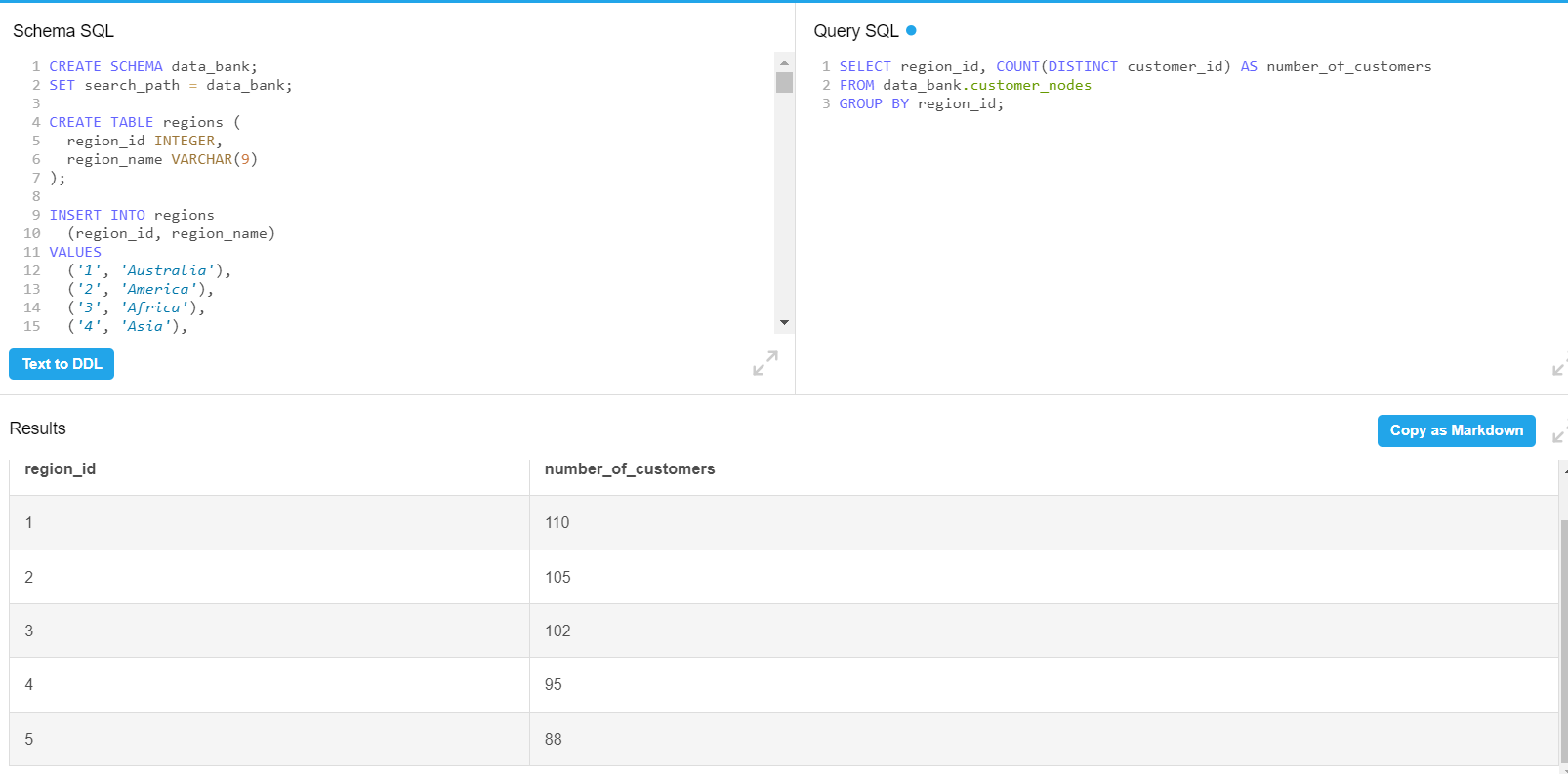
****

****

**RegionId 1 had the highest number of nodes occurrences (770), followed by 2 (735) with 5 having the least number of nodes (616).**

**3.How many customers are allocated to each region?**

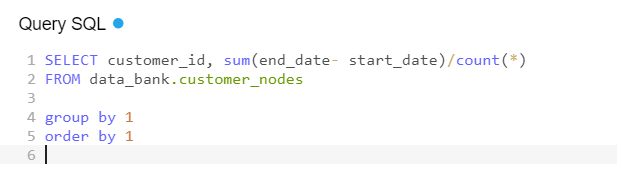
****

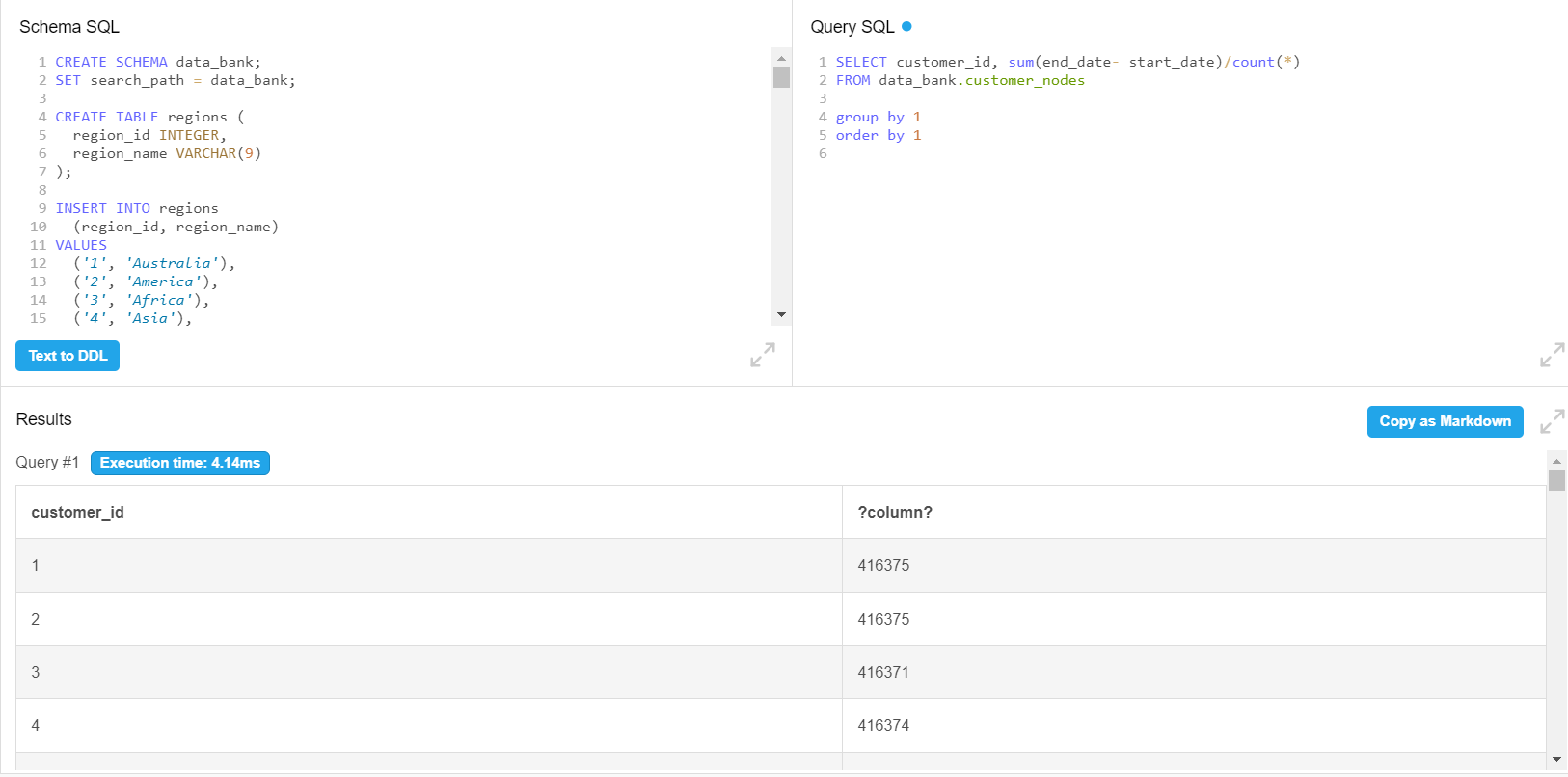


This query groups the data by region\_id and counts the **unique customers** (customer\_id) allocated to each region.

Region\_id 1 has the highest no. of customers followed by region\_id 2 and so on.

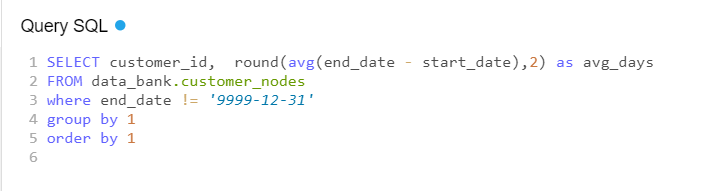
**4.How many days on average are customers reallocated to a different node?**

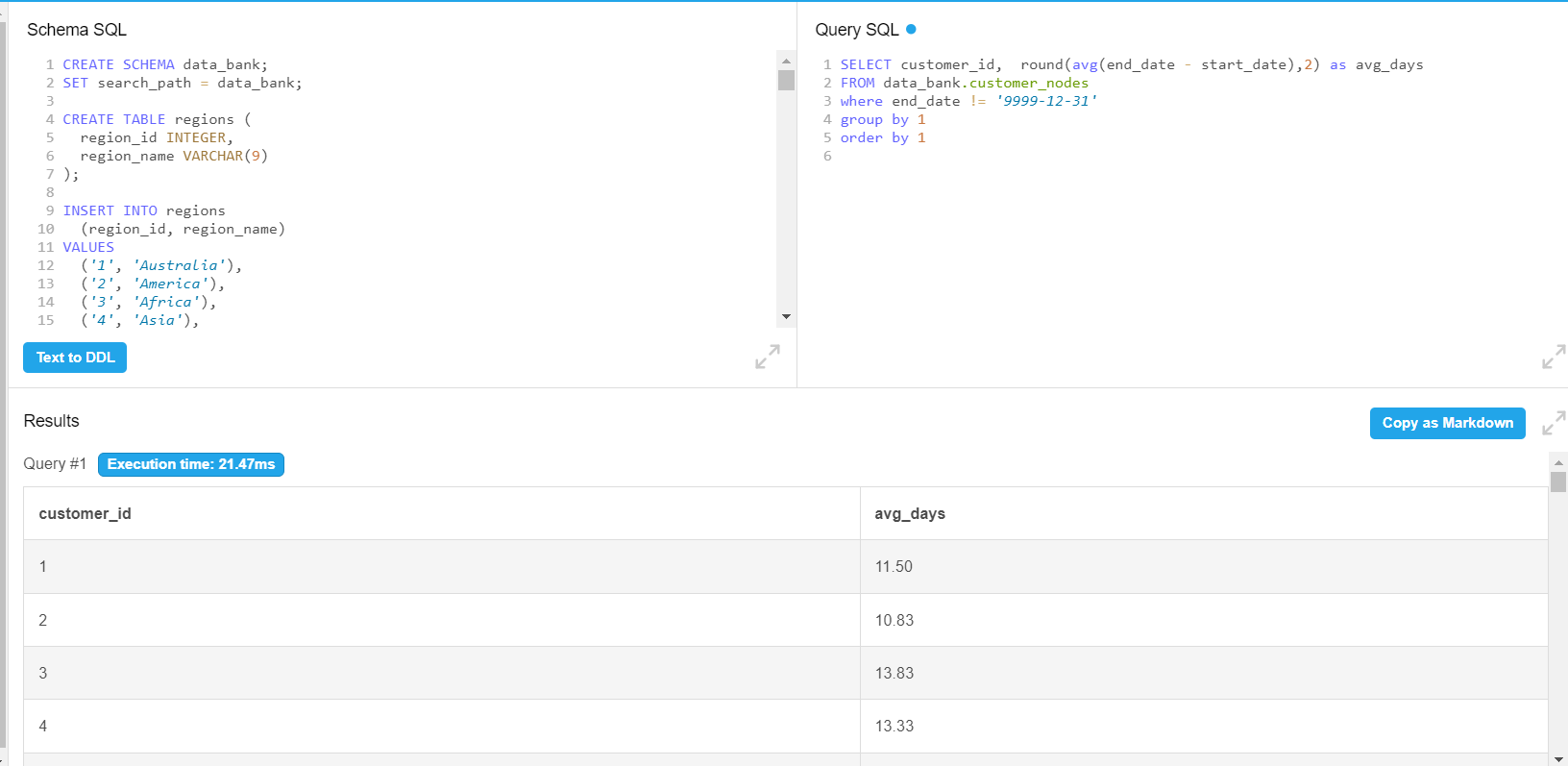


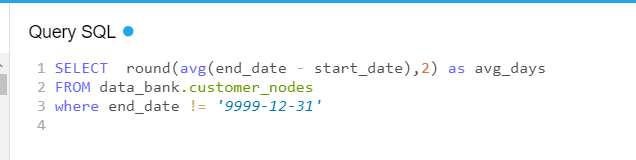


This query computes the **average number of days** customers stay allocated to a node by subtracting the start\_date from the end\_date and then averaging the results, **average number of days is 416373** but here is an outlier present in the data i.e end date **‘9999–12–31’** which is resulting in very highaverage no. of days and it can't be possible as per the security point of view of the data bank.

**Lets explore the same query by excluding the date ‘9999–12–31’.**

****

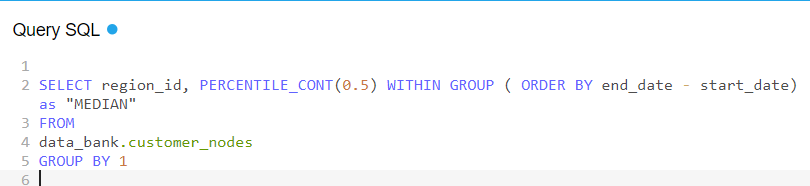
****

****

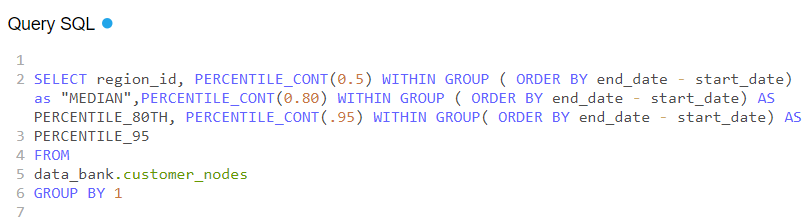
****

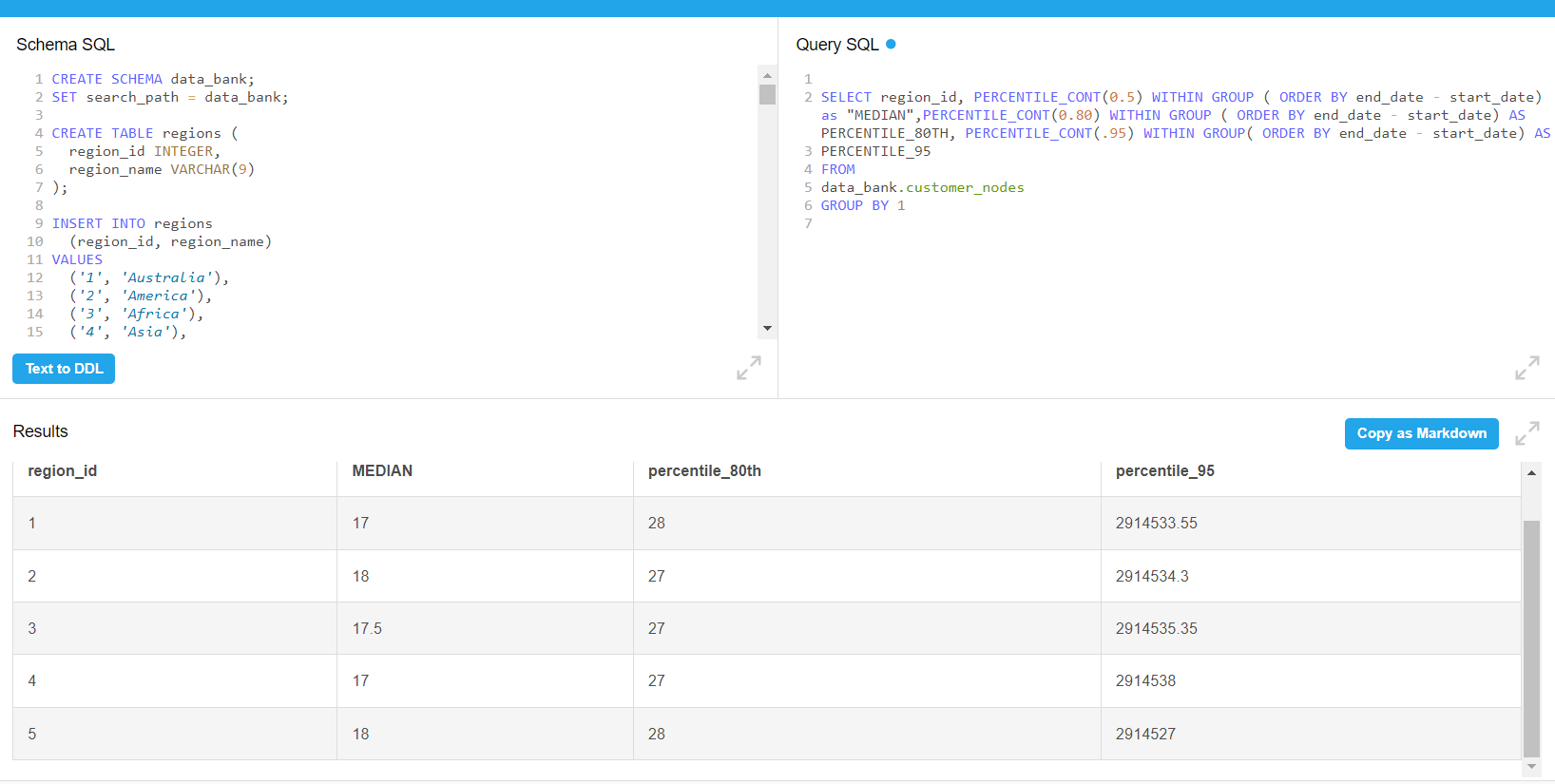
**It takes 14 days on average for customers to be reallocated to a different region.**

**5.What is the median, 80th and 95th percentile for this same reallocation days metric for each region?**







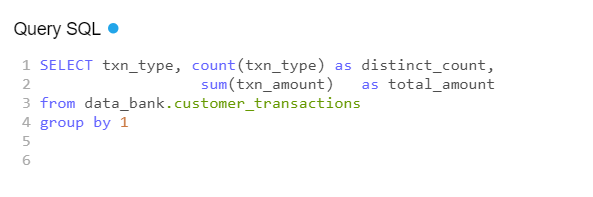


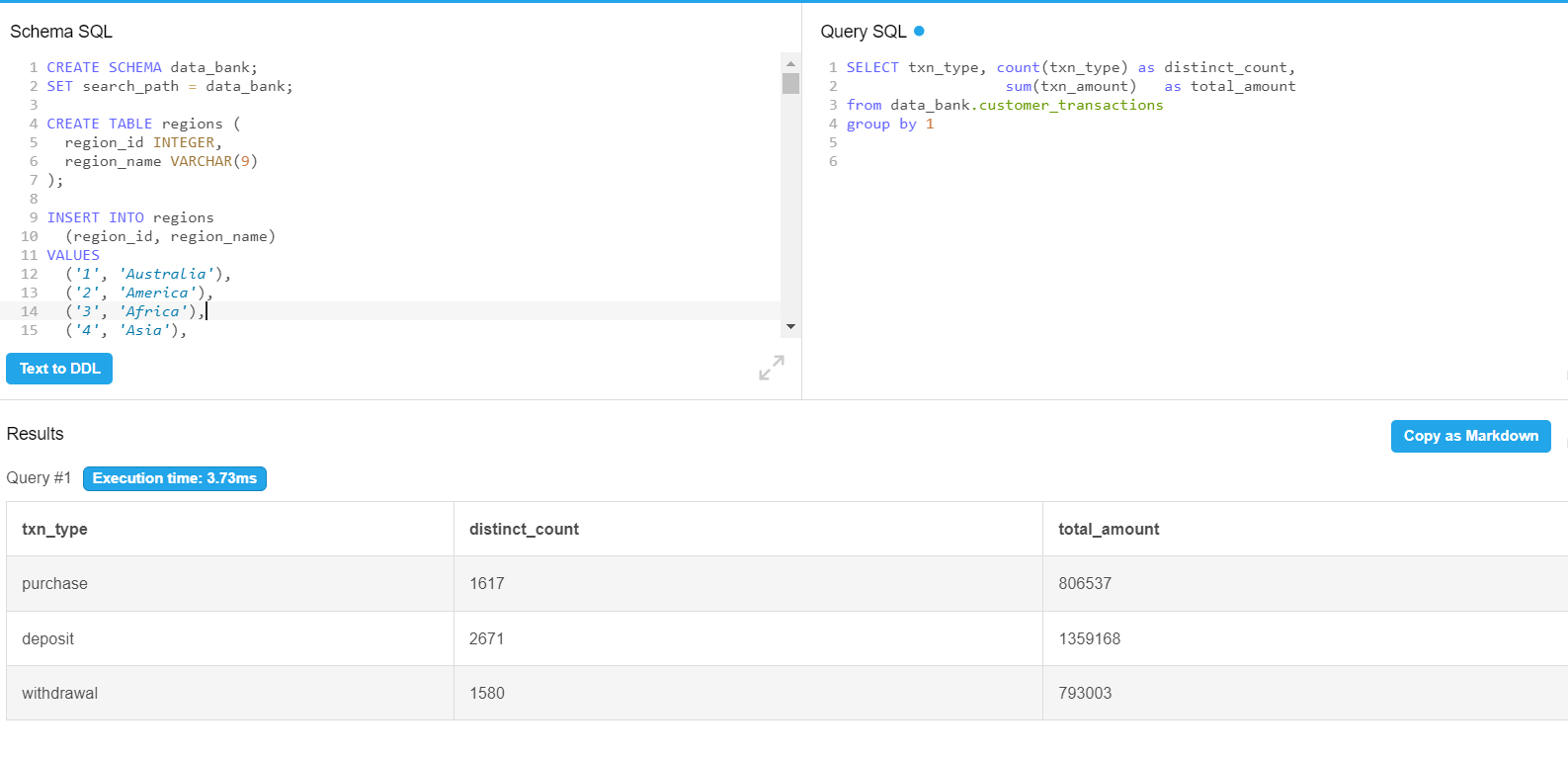
* **PERCENTILE\_CONT(0.50)** calculates the **median** (50th percentile).
* **PERCENTILE\_CONT(0.80)** calculates the **80th percentile**.
* **PERCENTILE\_CONT(0.95)** calculates the **95th percentile**.
* **end\_date - start\_date** calculates the number of days a customer is allocated to a node.
* The query is grouped by region\_id to get the statistics for each region.

This query will return the **median, 80th percentile, and 95th percentile** of reallocation days for each region.

B. Customer Transactions

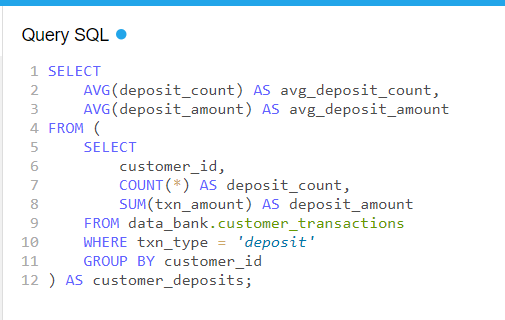
1. **What is the unique count and total amount for each transaction type?**

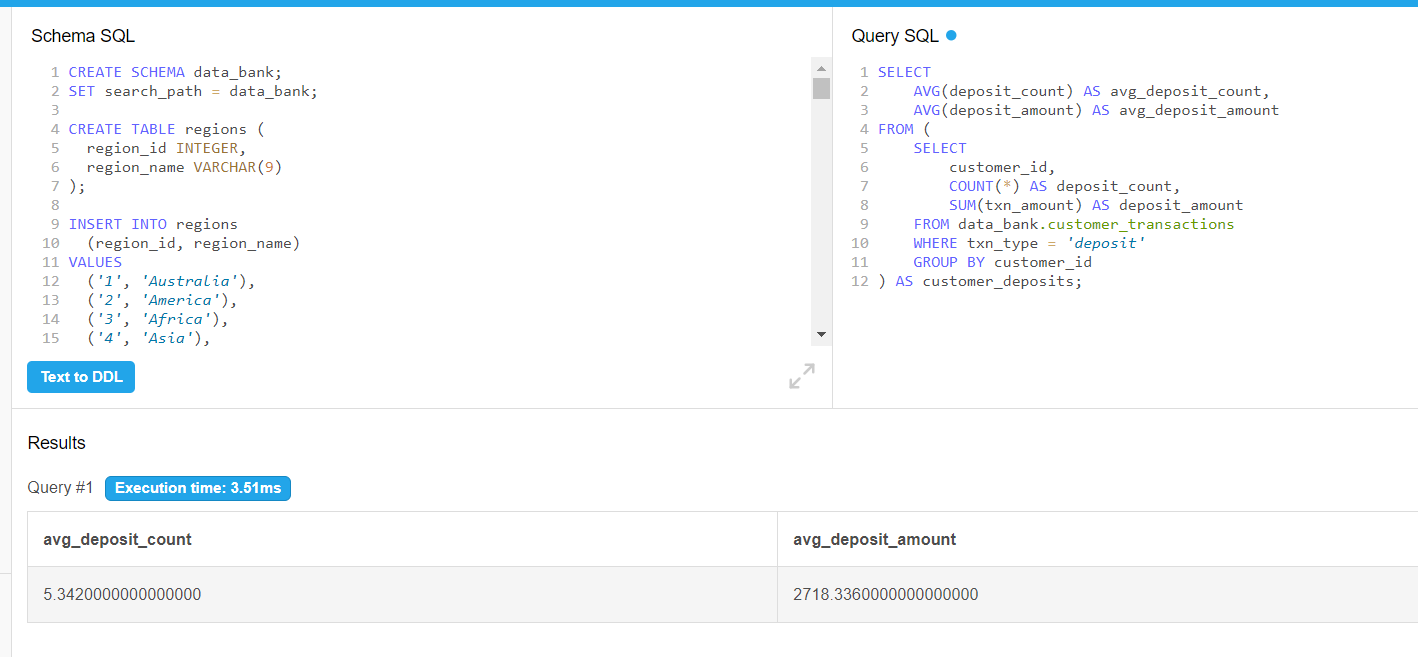
****

****

* **txn\_type**: Groups the data by each transaction type (e.g., deposit, withdrawal, purchase).
* **COUNT(DISTINCT txn\_type)**: Counts the number of unique transactions for each type.
* **SUM(txn\_amount):** Calculates the total transaction amount for each transaction type.
* **GROUP BY txn\_type**: Ensures the results are grouped by the different transaction types.
* There were more deposits (2671) followed by purchases(1617)and then withdrawals (1580).

**2.What is the average total historical deposit counts and amounts for all customers?**

****

****

Subquery (customer\_deposits):

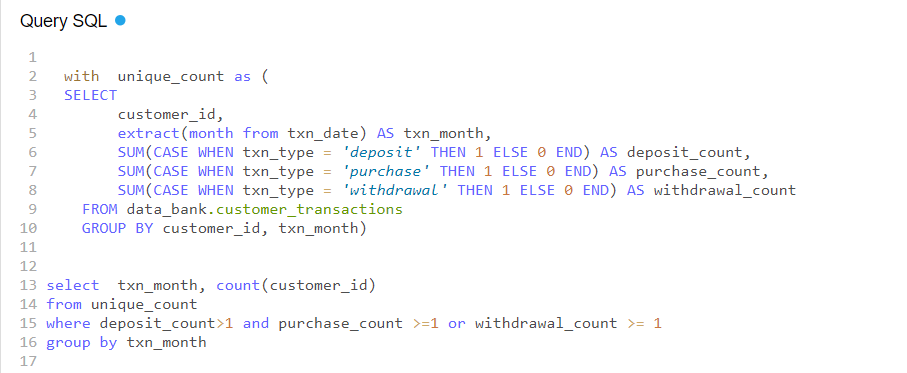
* The subquery filters for only deposit transactions (WHERE txn\_type = 'deposit').
* It groups by customer\_id to calculate:
  + COUNT(\*) AS deposit\_count: The total number of deposits per customer.
  + SUM(txn\_amount) AS deposit\_amount: The total deposit amount per customer.

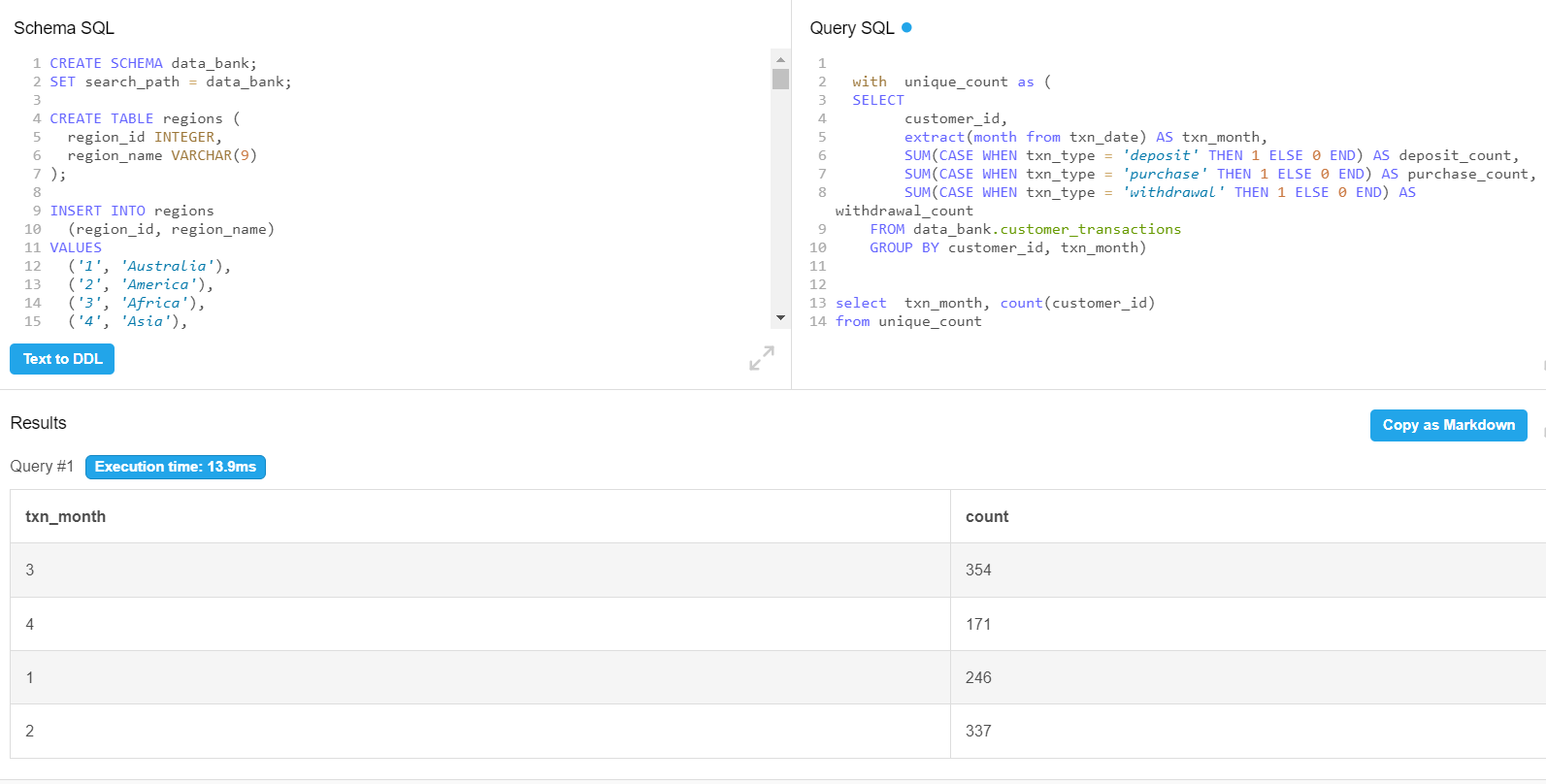
Main query:

* AVG(deposit\_count): Calculates the average deposit count across all customers.
* AVG(deposit\_amount): Calculates the average total deposit amount across all customers.

**The average deposit count for a customer is 5 and the average deposit amount for a customer is 2,718.**

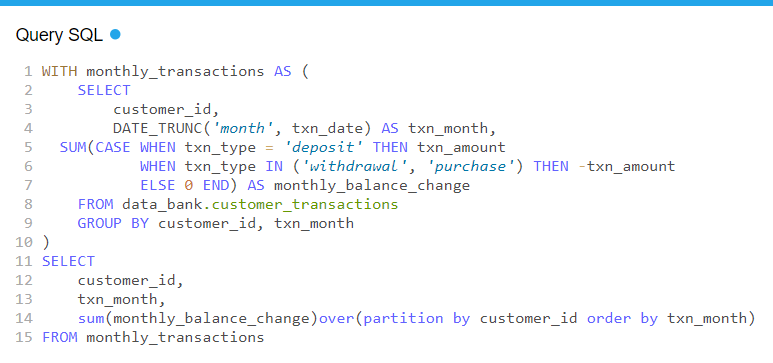
**3.For each month - how many Data Bank customers make more than 1 deposit and either 1 purchase or 1 withdrawal in a single month?**

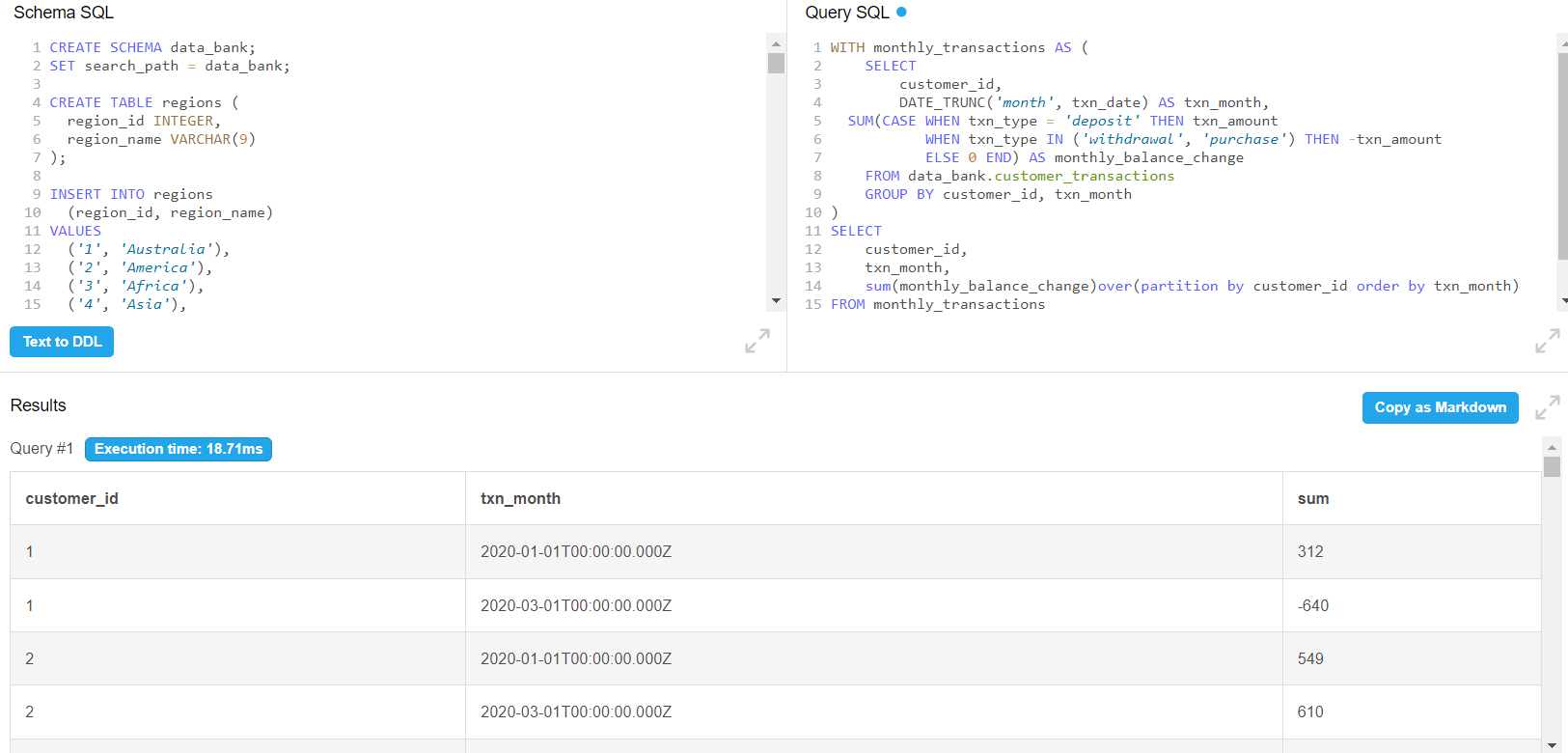
****

****

**March had the highest number of customers (354) who had made more than 1 deposit and either 1 withdrawal or 1 deposit while April had the least number of such customers (171).**

**4.What is the closing balance for each customer at the end of the month?**

****



**Monthly\_Transactions CTE**:

* **DATE\_TRUNC('month', txn\_date)**: Extracts the month from the transaction date to group transactions by month.
* **SUM(CASE...END)**: Calculates the net balance change for each customer in each month. Deposits increase the balance (txn\_amount is added), while withdrawals and purchases decrease it (txn\_amount is subtracted).

**Main query**:

* **SUM(...) OVER (PARTITION BY customer\_id ORDER BY txn\_month)**: This uses a window function to calculate the cumulative sum of the monthly balance changes for each customer. It tracks the running total (closing balance) at the end of each month.

Retrieves the **closing balance** for each customer at the end of each month, ordered by customer and month.

C. Data Allocation Challenge

To test out a few different hypotheses - the Data Bank team wants to run an experiment where different groups of customers would be allocated data using 3 different options:

**Option 1**: data is allocated based on the amount of money at the end of the previous month

**Option 2**: data is allocated on the average amount of money kept in the account in the previous 30 days

**Option 3**: data is updated real-time

For this multi-part challenge question - you have been requested to generate the following data elements to help the Data Bank team estimate how much data will need to be provisioned for each option:

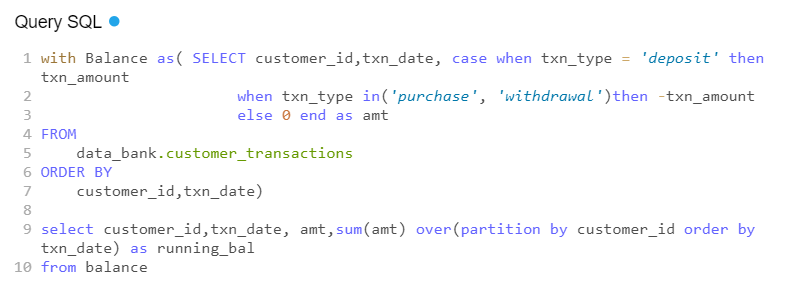
Using all of the data available - how much data would have been required for each option on a monthly basis?

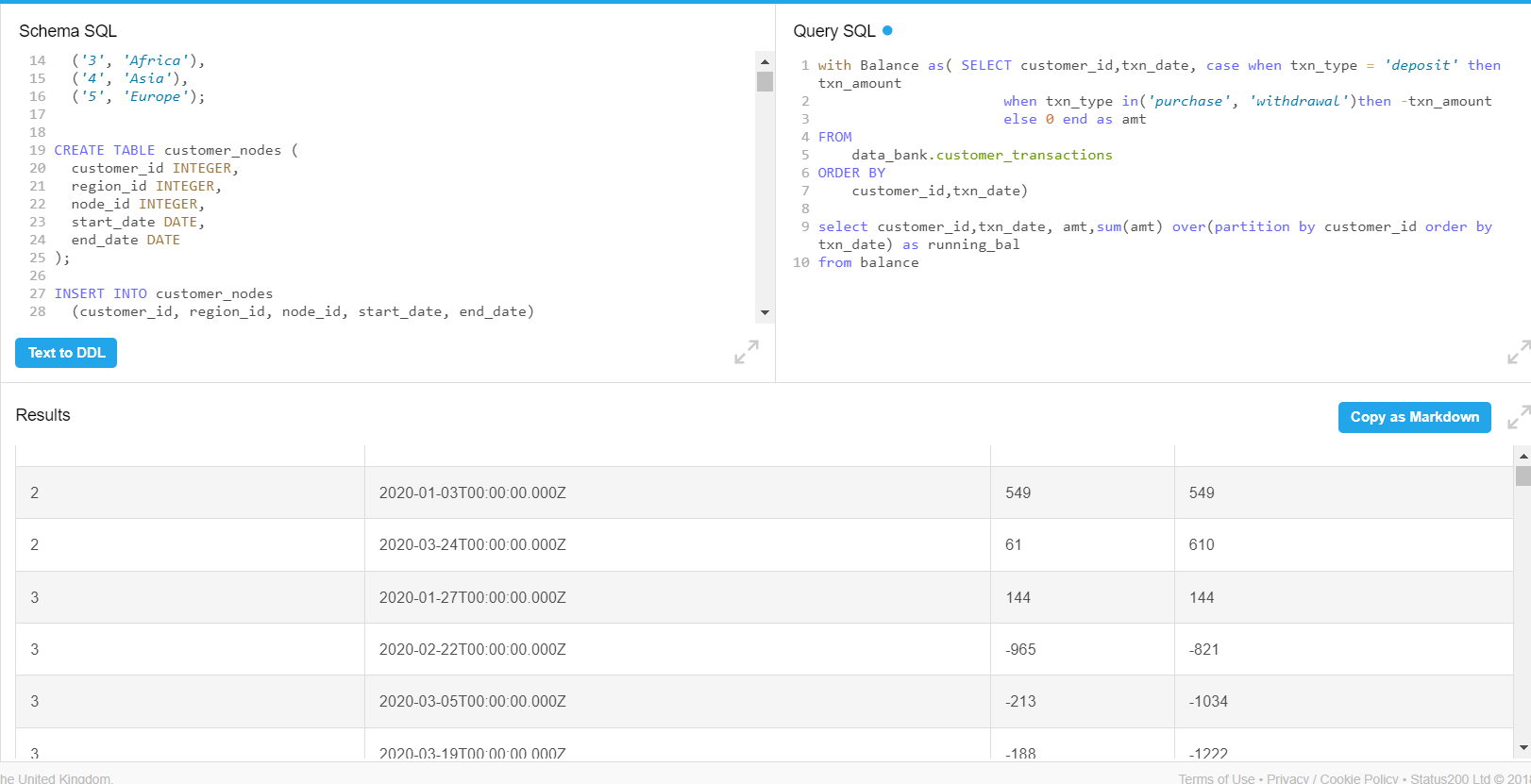
**1.Running customer balance column that includes the impact each transaction**

**Steps:**

Calculate the running balance for each customer based on the order of their transaction.

Adjust the 'txn\_amount' to be negative for withdrawal and purchase transactions to reflect a negative balance.



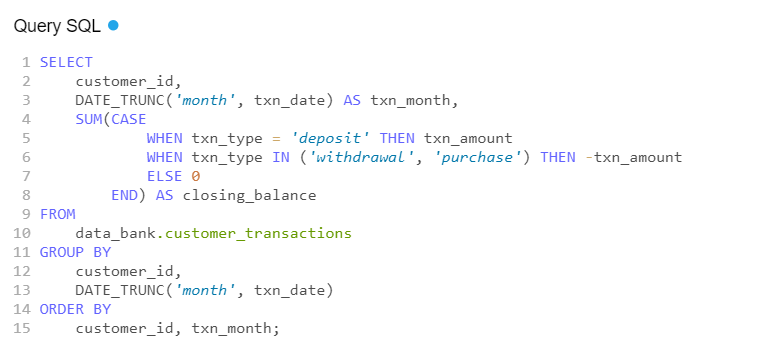


*Kindly note that this is not the entire output. The entire output is long and would take up space.*

**2.Customer balance at the end of each month**

**Steps:**

* Calculate the running balance for each customer based on the order of their transaction.
* Adjust the 'txn\_amount' to be negative for withdrawal and purchase transactions to reflect a negative balance.

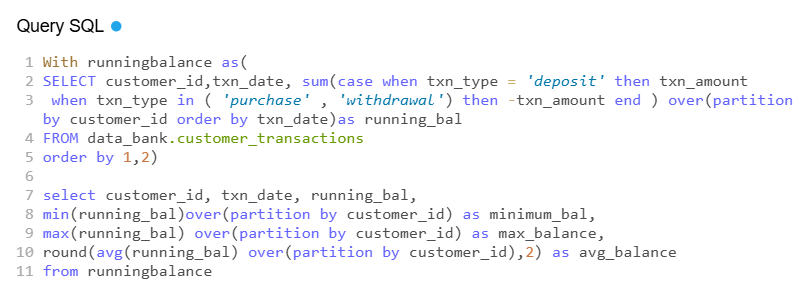




**3.Minimum, average and maximum values of the running balance for each customer**

**Steps:**

* Adjust the 'txn\_amount' to be negative for withdrawal and purchase transactions to reflect a negative balance.
* Calculate the running balance for each customer based on the order of their transaction.
* Extracted minimum, maximum, average of the running balance for each customer.

****



Now based on the provided three options, we will use each of the above calculated data elements to calculate how much data would have been required for each data allocation option on a monthly basis.

**For Option 1: Data is allocated based on the amount of money at the end of the previous month. How much data would have been required on a monthly basis?**

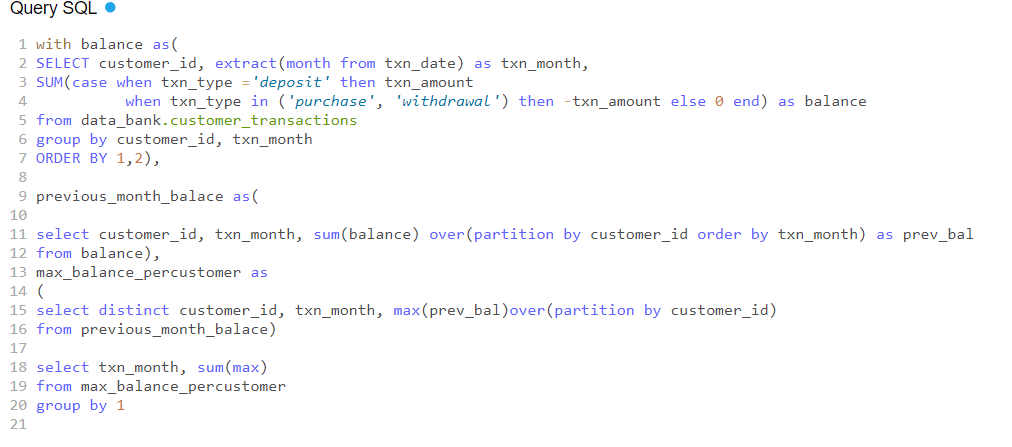
Steps:

Use a CTE to calculate the customer balance of each customer at the end of the month, separately month wise.

Use a CTE to calculate the customer balance at the end of every month using running total month wise

Use a third CTE to calculate maximum balance kept by customer across all the months.

Use the final query to calculate the data required per month by summing up the maximum balances kept by each customer.

****



**Insghts:**

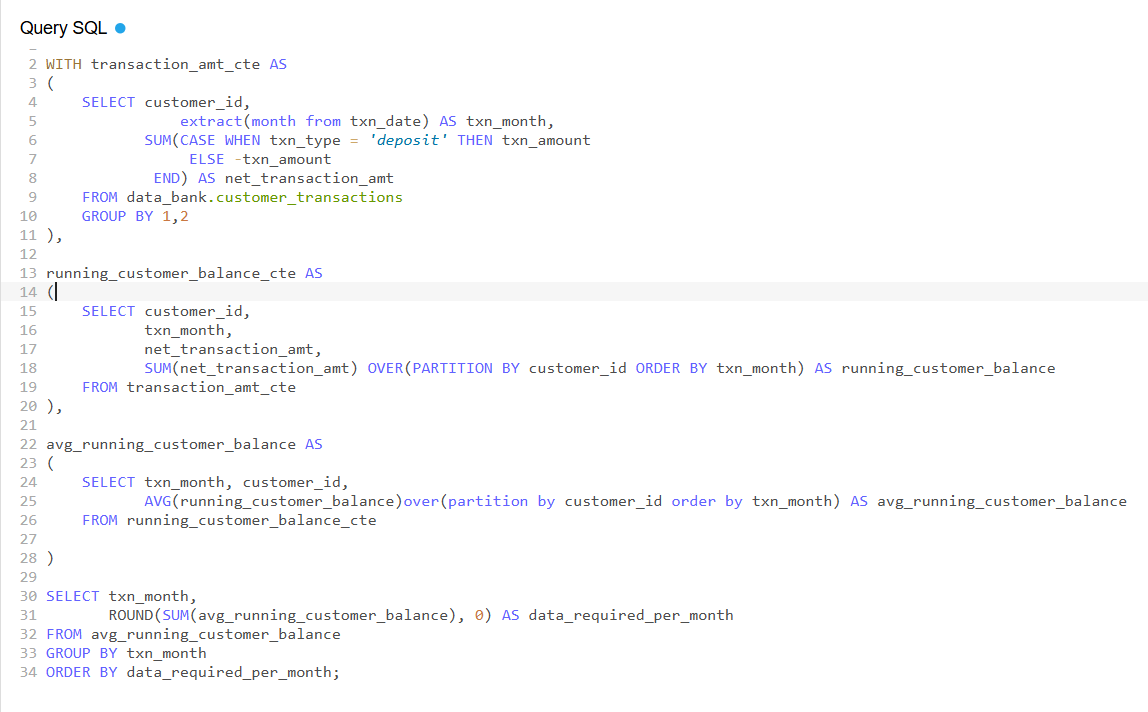
* January requires more monthly data allocation (377318) followed by March (347574), and February (342984), with April (248432) requiring the least amount of data.
* This actually means that data allocation that would be required per month varies across different months.
* This insight suggests that the amount of data required by customers is directly related to their transaction activities, and specifically to their end-of-month balances. This indicates that customers with higher balances tend to require more data than those with lower balances.
* In other words, customers tend to do have higher end month balances in January and March than in Frebruary and April so more data should be allocated for January, followed by March, February and April.
* This insight generated would help predict customer behaviour, optimizing business strategies and managing costs.

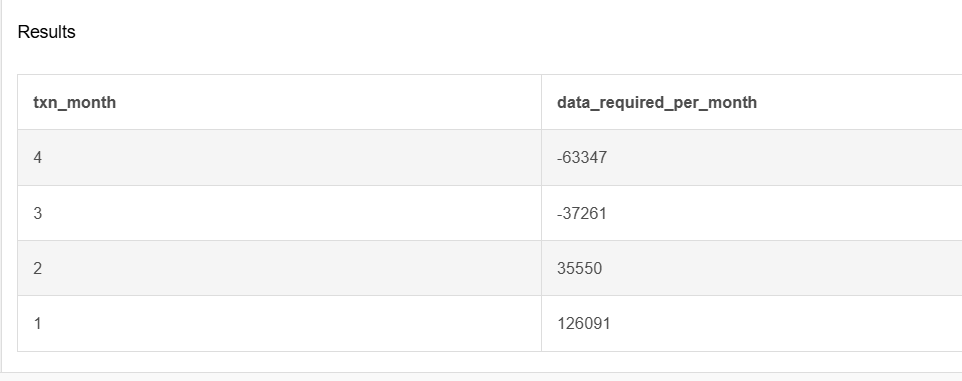
**Option 2: data is allocated on the average amount of money kept in the account in the previous 30 days.**

Steps:

* Use transaction\_amt\_cte CTE to calculate the net transaction amount for each customer in each month.
* Use running\_customer\_balance\_cte CTE to calculate the running balance for each customer in each month, based on the net transaction amount
* Use avg\_running\_customer\_balance CTE to calculate the average running customer balance for ecah customer across all months.
* In the final query, group the data by month, and calculate the rounded sum of the average running customer balance as data\_required\_per\_month.

This gives an estimate of how much data would be required for option 2 on a monthly basis.

****



**Insights:**

* Based on our query output, the average running customer balance is negative for march and April months, indicating that customers tend to withdraw more money than they deposit on average.
* The data required for January and February are higher, suggesting that more data should be allocated for those two months.
* These negative running balances, could impact the bank's overall financial health. Therefore, I recommend that the bank collect more data for April and March to better understand customer behaviour during those months and potentially identify any trends or anomalies that could impact the bank's business.

**Option 3: data is updated real-time.**

**How much data would have been required on a monthly basis?**

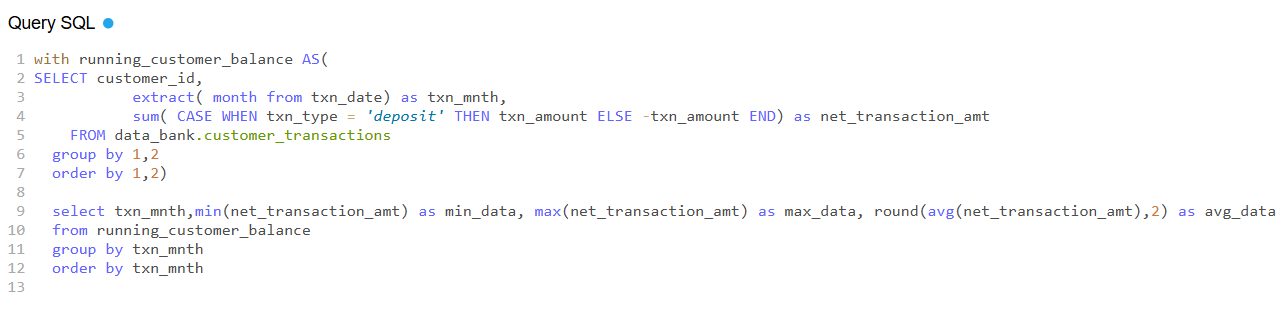
Steps:

* Use running\_customer\_balance CTE to calculate the running balance for each customer by summing up the net transaction amounts over time(months).
* Use the final query to calculate the estimated data required per month for option 3, assuming

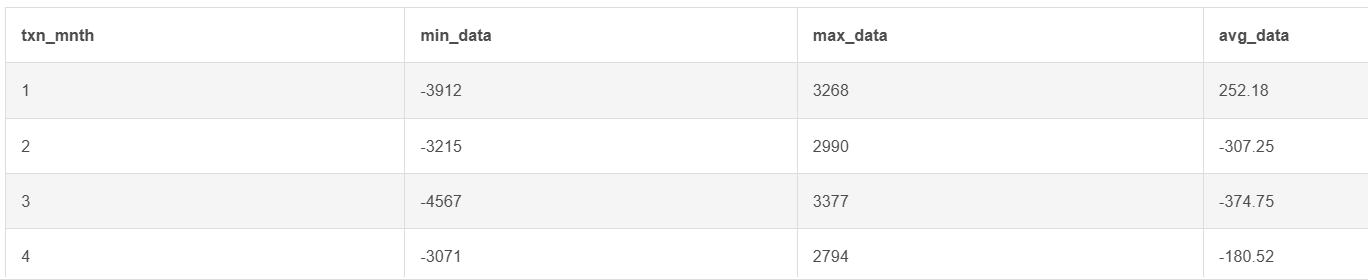
maximum, minimum, and average amount of money for each month across all the customers

to predict the real-time data requirement.

* Assumption The amount of money kept in the account is considered as the amount of data required.



Result:



**Insights**

* The data required for the month of March is significantly higher than for the other months. This shows that there were more transactions happening in March than in the other months.
* Avg\_data required for January is positive, indicating that there might be some customers who have a higher balance at the beginning of the year.
* it might possible some of the customers keep low balance over all the months and because of this the minimum data required is always negative.

**Recommendation**

* Data Bank should bring out a clause to maintain the minimum average monthly balance to keep the minimum balance high.

D. Extension Request

The Data Bank team wants you to use the outputs generated from the above sections to create a quick Power point presentation which will be used as marketing materials for both external investors who might want to buy Data Bank shares and new prospective customers who might want to bank with Data Bank.

Using the outputs generated from the customer node questions, generate a few headline insights which Data Bank might use to market it’s world-leading security features to potential investors and customers.

With the transaction analysis – prepare a presentation slide which contains all the relevant information about the various options for the data provisioning so the Data Bank management team can make an informed decision.

Here's a[Power Point presentation](https://www.canva.com/design/DAGTu7ybR1Y/3tT8mo8PHWXYS6A_SDp_lA/edit?utm_content=DAGTu7ybR1Y&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton) that meets the requirements of the above request.