

# SQL CASE STUDY ON DATA BANK

Table 1: Regions

region_id	region_name
1	Africa
2	America
3	Asia
4	Europe
5	Oceania

Table 2: 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**

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

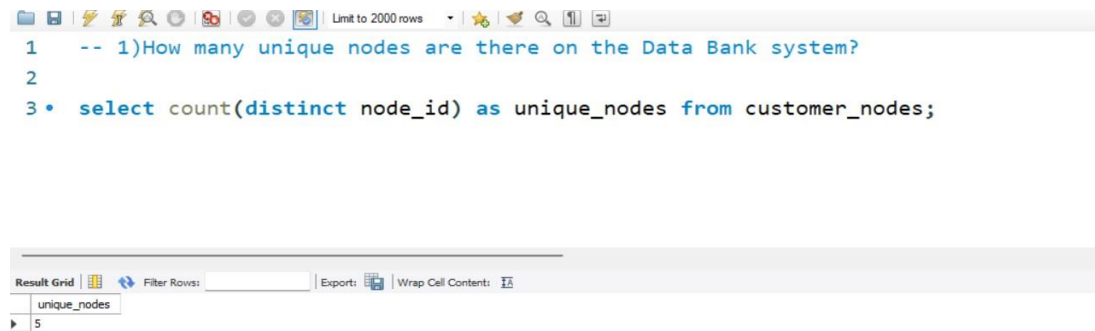
### **A. Customer Nodes Exploration**

1. How many unique nodes are there on the Data Bank system?
2. What is the number of nodes per region?
3. How many customers are allocated to each region?
4. How many days on average are customers reallocated to a different node?
5. What is the median, 80th and 95th percentile for this same reallocation days metric for each region?

### **B. Customer Transactions**

1. What is the unique count and total amount for each transaction type?
2. What is the average total historical deposit counts and amounts for all customers?
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?

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

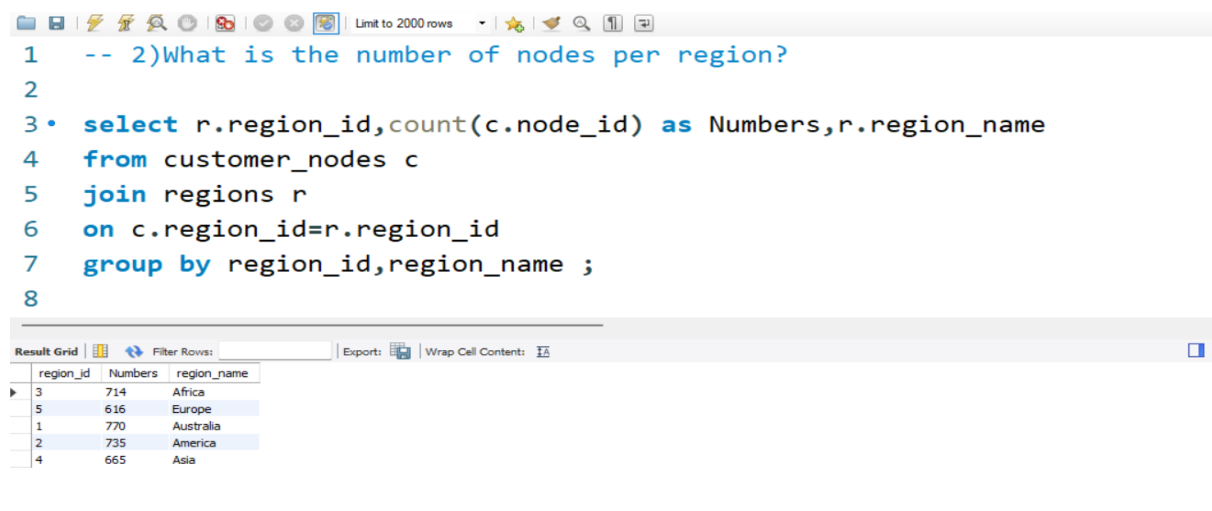


The screenshot shows a SQL IDE interface with a toolbar at the top. Below the toolbar, a SQL query is entered in a text area. The query is: `-- 1)How many unique nodes are there on the Data Bank system?`  
`2`  
`3 • select count(distinct node_id) as unique_nodes from customer_nodes;`

Below the query editor, the 'Result Grid' is visible. It shows a single row with the column name 'unique\_nodes' and the value '5'.

unique_nodes
5

## 2)What is the number of nodes per region?

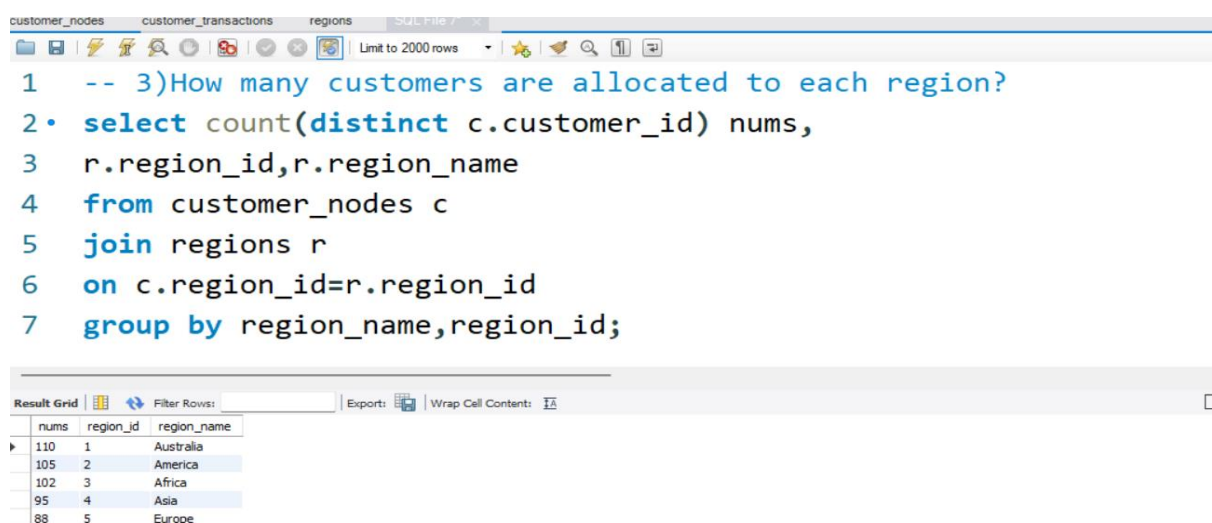


The screenshot shows a SQL IDE interface with a toolbar at the top. Below the toolbar, a SQL query is entered in a text area. The query is: `-- 2)What is the number of nodes per region?`  
`2`  
`3 • select r.region_id,count(c.node_id) as Numbers,r.region_name`  
`4 from customer_nodes c`  
`5 join regions r`  
`6 on c.region_id=r.region_id`  
`7 group by region_id,region_name ;`  
`8`

Below the query editor, the 'Result Grid' is visible. It shows a table with three columns: 'region\_id', 'Numbers', and 'region\_name'. The data is as follows:

region_id	Numbers	region_name
3	714	Africa
5	616	Europe
1	770	Australia
2	735	America
4	665	Asia

## 3)How many customers are allocated to each region?

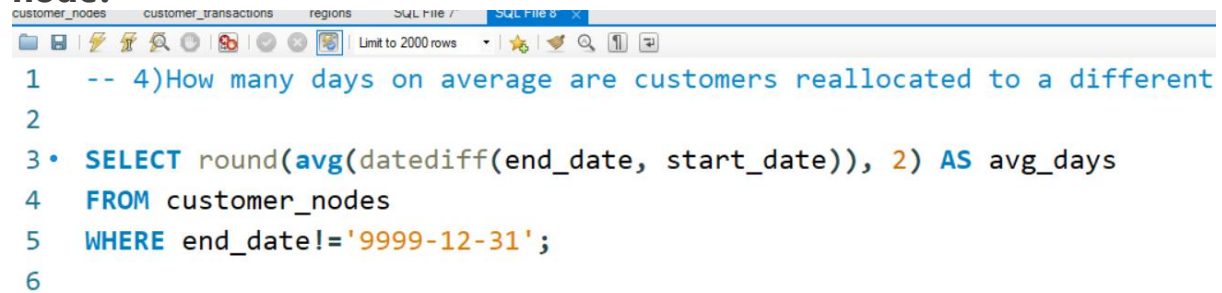


The screenshot shows a SQL IDE interface with a toolbar at the top. Below the toolbar, a SQL query is entered in a text area. The query is: `-- 3)How many customers are allocated to each region?`  
`2 • select count(distinct c.customer_id) nums,`  
`3 r.region_id,r.region_name`  
`4 from customer_nodes c`  
`5 join regions r`  
`6 on c.region_id=r.region_id`  
`7 group by region_name,region_id;`

Below the query editor, the 'Result Grid' is visible. It shows a table with three columns: 'nums', 'region\_id', and 'region\_name'. The data is as follows:

nums	region_id	region_name
110	1	Australia
105	2	America
102	3	Africa
95	4	Asia
88	5	Europe

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



```
1  -- 4)How many days on average are customers reallocated to a different
2
3 • SELECT round(avg(datediff(end_date, start_date)), 2) AS avg_days
4 FROM customer_nodes
5 WHERE end_date != '9999-12-31';
6
```



avg_days
14.63

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

```
WITH reallocation_days_cte AS (
    SELECT *,
        (datediff(end_date, start_date)) AS reallocation_days
    FROM customer_nodes
    INNER JOIN regions USING (region_id)
    WHERE end_date != '9999-12-31'
),
percentile_cte AS (
    SELECT *,
        percent_rank() OVER (PARTITION BY region_id ORDER BY reallocation_days) * 100 AS p
    FROM reallocation_days_cte
)
SELECT region_id,
    region_name,
    reallocation_days
FROM percentile_cte
WHERE p > 80
group by region_id,region_name,reallocation_days
;
```

## B. Customer Transactions

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

```
customer_nodes customer_transactions regions SQL File 12" x
Limit to 2000 rows
1 -- 1) What is the unique count and total amount for each transaction type?
2 • select count(distinct customer_id) as dist_count,
3     sum(txn_amount) as total,
4     txn_type
5 from customer_transactions
6 group by txn_type
7 ;
```

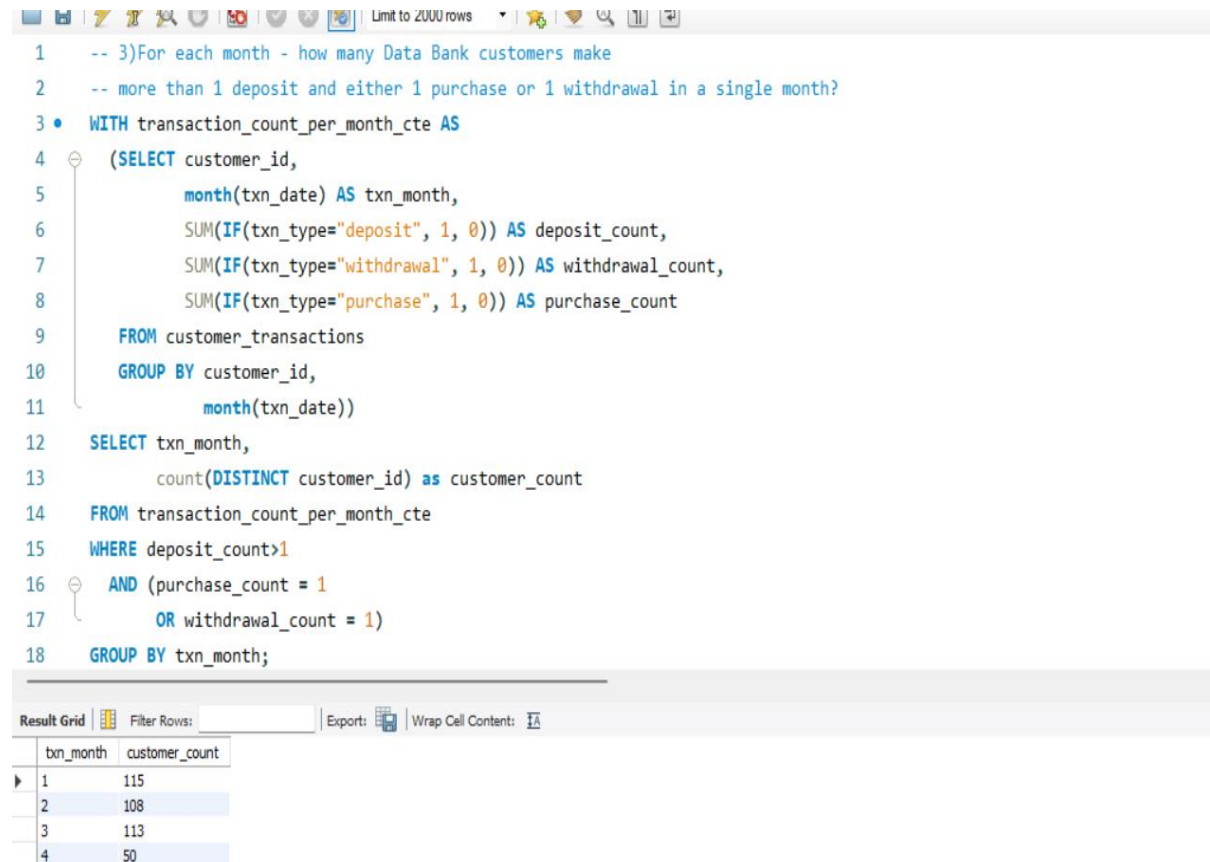
Result Grid	Filter Rows:	Export:	Wrap Cell Content:
dist_count	total	txn_type	
500	1359168	deposit	
448	806537	purchase	
439	793003	withdrawal	

### 2) What is the average total historical deposit counts and amounts for all customers?

```
Limit to 2000 rows
1 -- 2) What is the average total historical deposit counts and amounts for all customers?
2 • select avg(deposit_counts) as avg_deposit_counts,
3     avg(total) as avg_total_transcation
4 from
5 (
6     select customer_id,txn_type,
7     count(txn_type) as deposit_counts,
8     sum(txn_amount) as total from
9     customer_transactions
10    where txn_type='deposit'
11    group by customer_id
12 )
13 as summery;
```

Result Grid	Filter Rows:	Export:	Wrap Cell Content:
avg_deposit_counts	avg_total_transcation		
5.3420	2718.3360		

### 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?



The screenshot shows a SQL IDE interface. The top toolbar includes icons for file operations, a search icon, and a 'Limit to 2000 rows' dropdown. The SQL editor contains a query with line numbers 1 through 18. The query defines a CTE named 'transaction\_count\_per\_month\_cte' and then selects from it. The results pane at the bottom shows a table with two columns: 'txn\_month' and 'customer\_count', with four rows of data.

```
1 -- 3)For each month - how many Data Bank customers make
2 -- more than 1 deposit and either 1 purchase or 1 withdrawal in a single month?
3 • WITH transaction_count_per_month_cte AS
4   (SELECT customer_id,
5           month(txn_date) AS txn_month,
6           SUM(IF(txn_type="deposit", 1, 0)) AS deposit_count,
7           SUM(IF(txn_type="withdrawal", 1, 0)) AS withdrawal_count,
8           SUM(IF(txn_type="purchase", 1, 0)) AS purchase_count
9        FROM customer_transactions
10       GROUP BY customer_id,
11                month(txn_date))
12 SELECT txn_month,
13        count(DISTINCT customer_id) as customer_count
14 FROM transaction_count_per_month_cte
15 WHERE deposit_count > 1
16    AND (purchase_count = 1
17         OR withdrawal_count = 1)
18 GROUP BY txn_month;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: |

	txn_month	customer_count
▶	1	115
	2	108
	3	113
	4	50