Data Bank SQL Challenge - PySpark Solution Documentation

Case Study Link 🡪 <https://8weeksqlchallenge.com/case-study-4/>

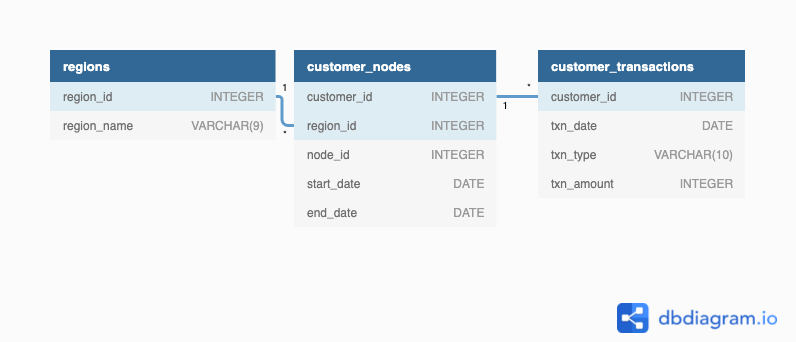
Databricks Notebook 🡪 <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2849348766235554/1518048595388115/7676470150836731/latest.html>

Git Repo 🡪 <https://github.com/Nasreen-Git-Repo/Danny-Ma-s-8-Week-SQL-Challenge>



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!

**Entity Relationship Diagram**



**Table 1: Regions**

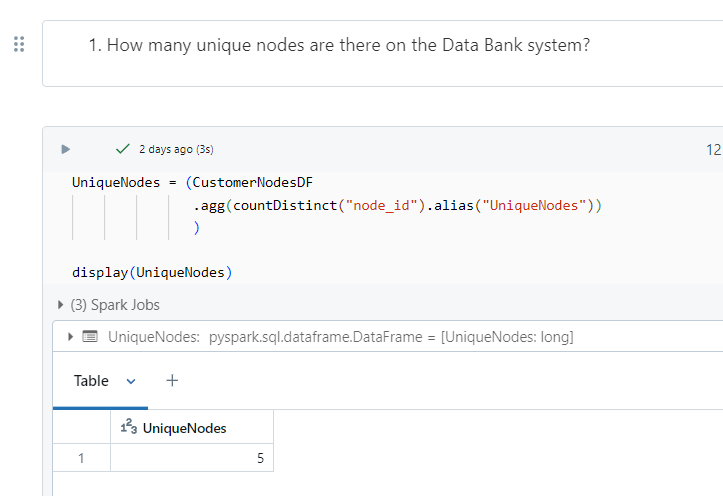
**Table 2: Customer Nodes**

**Table 3: Customer Transactions**

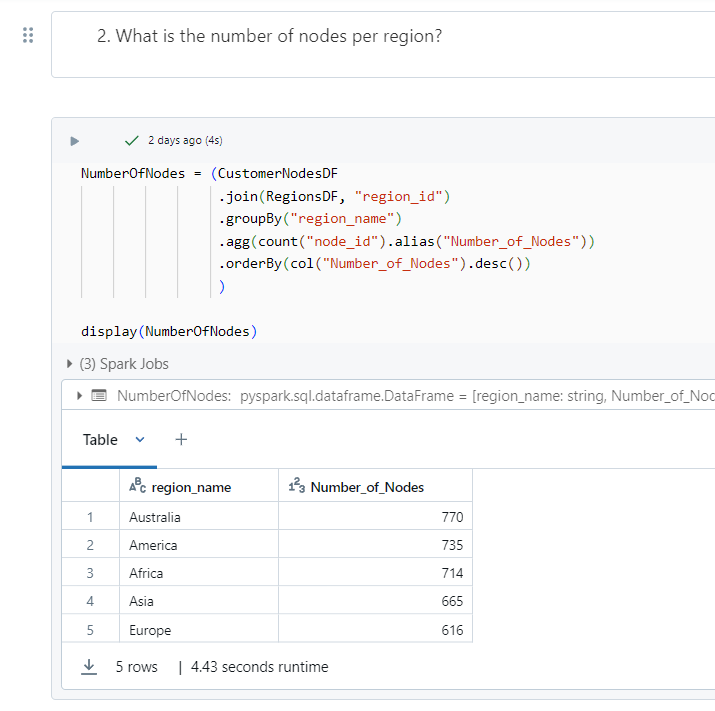
The Exact dataset is in Databricks Notebook as well as on Case Study Page

**A. Customer Nodes Exploration**

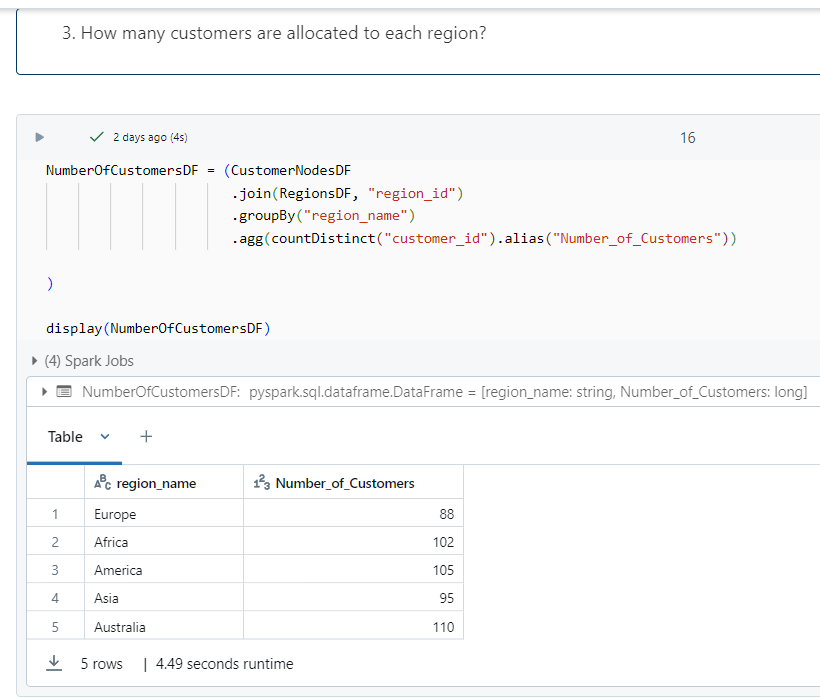
1. How many unique nodes are there on the Data Bank system?

****

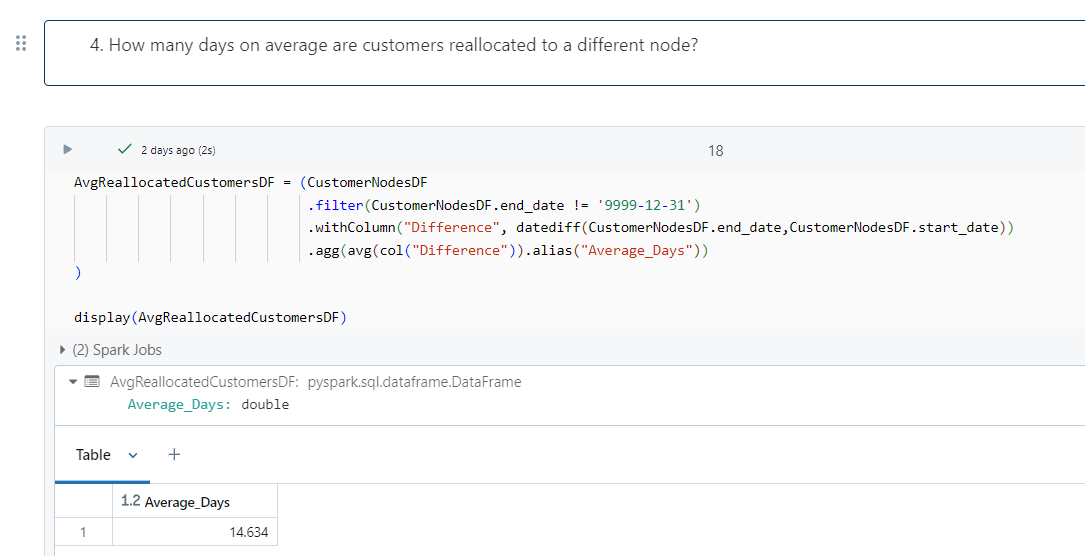
1. What is the number of nodes per region?



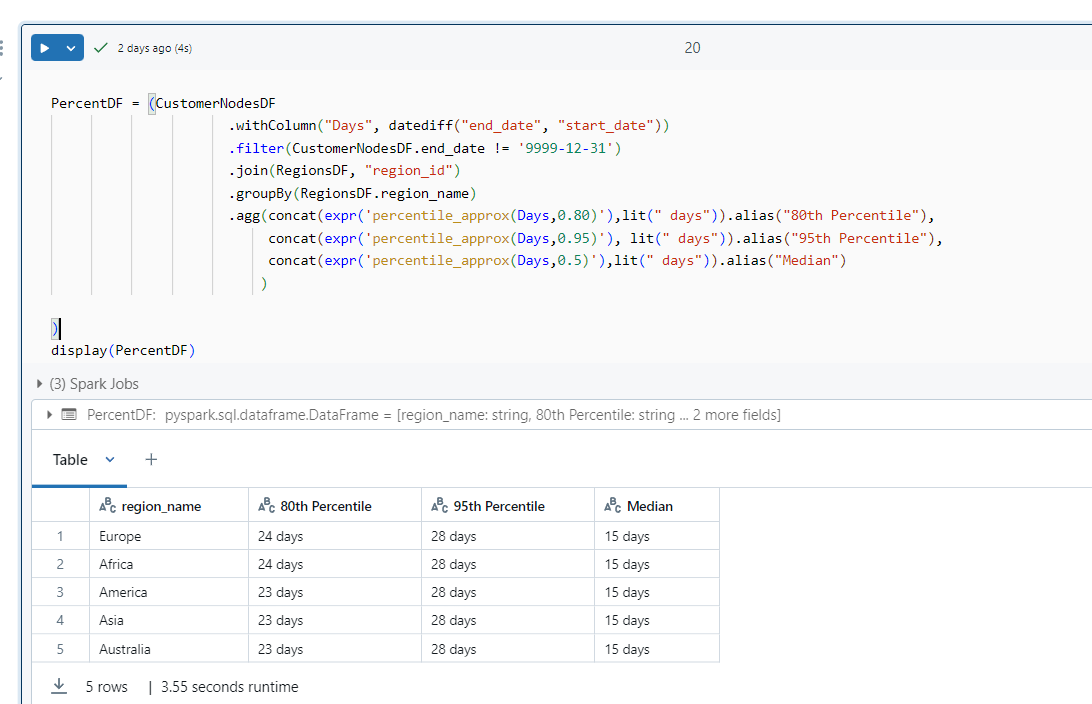
1. How many customers are allocated to each region?



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

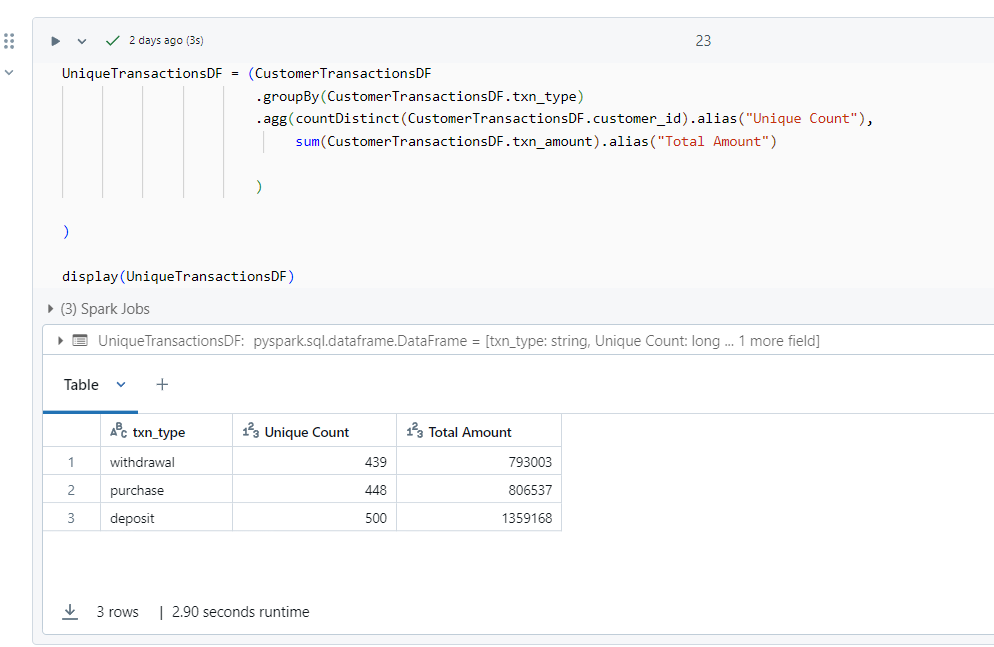


1. What is the median, 80th and 95th percentile for this same reallocation days metric for each region?



**B. Customer Transactions**

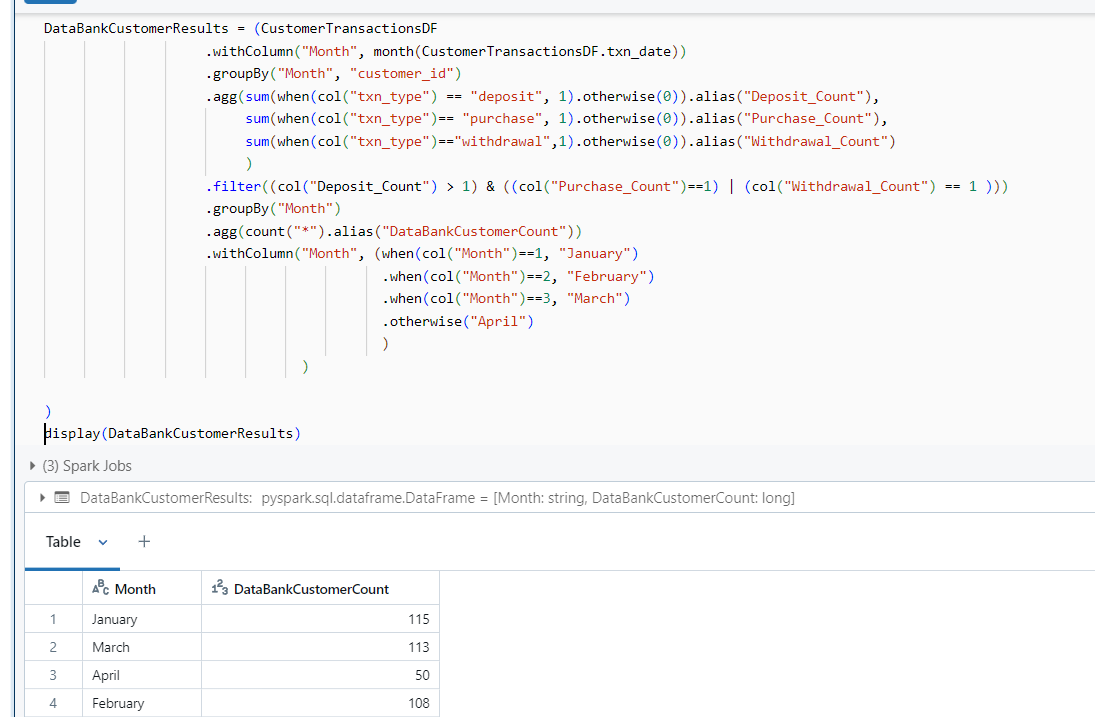
1. What is the unique count and total amount for each transaction type?



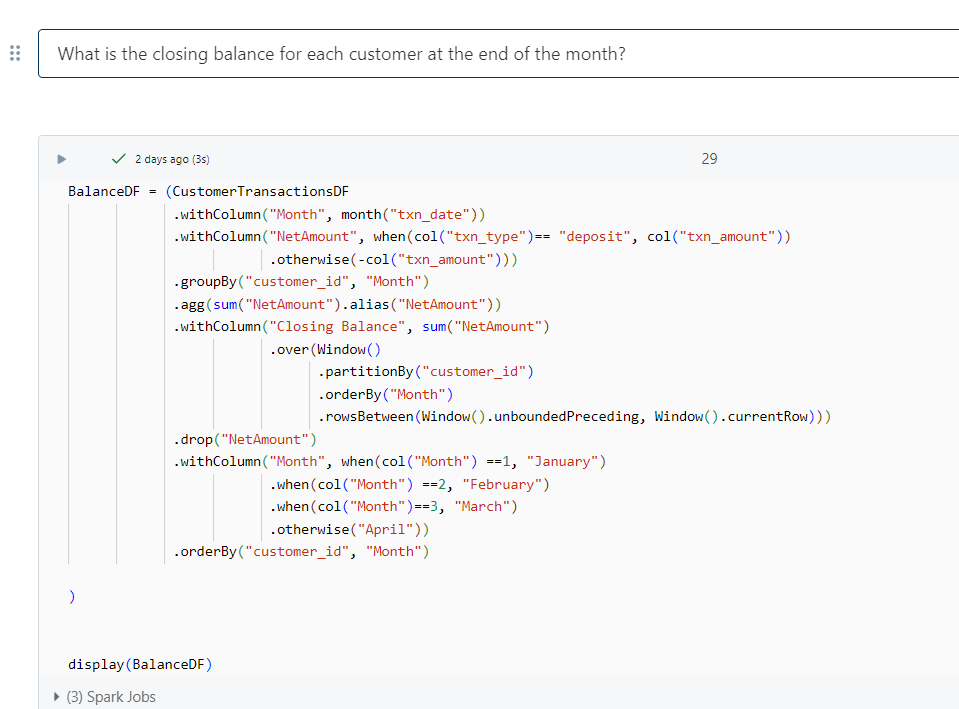
1. What is the average total historical deposit counts and amounts for all customers?



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



It is not feasible to copy entire output here. Below is the sample output

