Juliana Porto

February 10, 2024

DS503 Big Data Management

Project 1

Question	**Status**	**Comment**
Q1	Fully Working	I created 2 datasets as outlined in the Project 1 instructions with the customers in Customers.txt and the transactions in Transactions.txt. **Source Code:** Project1/Q1_CreatingDatasets/Main.java **Customers Dataset:** Project1/Q1_CreatingDatasets/Customers.txt **Transactions Dataset:** Project1/Q1_CreatingDatasets/Transactions.txt
Q2	Fully Working	Location of customer dataset: /user/Project1/data/Customers.txt Location of transaction dataset: /user/Project1/data/Transactions.txt I also checked the files to see how the files are divided into blocks and how each block is replicated. I found that Customers.txt was located within 1 block, but Transactions.txt was located within 3 blocks. **Screenshot of HDFS:** Project1/Q2_UploadingDataIntoHadoop/HDFS_data.png
Q3.1	Fully Working	*Query Plan:* Customer Map Logic: 1. For a given record, do parsing and extract fields 2. Key = ID value 3. Value = "Customer", ID value, Name value, Salary value Transaction Map Logic: 1. For a given record, do parsing and extract fields 2. Key = CustID value

		3. Value = “Transaction”, TransTotal, TransNumItems Reduce Logic: 1. Separate array based on dataset source, either “Customer” or “Transaction” 2. Join records based on ID and CustID 3. For the joined tuples, apply the count aggregation function to TransTotal to get NumOfTransactions 4. For the joined tuples, apply the sum aggregation function to TransTotal to get TotalSum 5. For the joined tuples, apply the min aggregation function to TransNumItems to get MinItems 6. Output: key = NULL; value = CustID, Name, Salary, NumOfTransactions, TotalSum, MinItems **Source Code:** Project1/Q3_WritingMapReduceJobs/Q3.1_Query1/CustomerTransactionJoin.java **Output:** Project1/Q3_WritingMapReduceJobs/Q3.1_Query1/FinalOutputQ3a/part-r-00000
Q3.2	Fully Working	**Note**: Took a few solid attempts at completing this query with just one job but am unsure of how it would work since the join key does not equal the group by key as explained in the slides. I understand that I will lose 8 points because of this, but the code is technically fully working :) *Query Plan:* Customer Map Logic: 1. For a given record, do parsing and extract fields 2. Key = ID value 3. Value = “Customer”, ID value, CountryCode value Transaction Map Logic: 1. For a given record, do parsing and extract fields 2. Key = CustID value 3. Value = “Transaction”, TransTotal Reduce Logic: 1. Separate array based on dataset source, either “Customer” or “Transaction” 2. Join records

		3. Output: key = NULL; value = ID, CountryCode, TransTotal Map 2 Logic: 1. For a given record, do parsing and extract fields 2. Key = CountryCode value 3. Value = CountryCode, TransTotal Reduce 2 Logic: 1. Set CountryCode and TransTotal 2. Set minTransTotal 3. Set maxTransTotal 4. Increment numberOfCustomers 5. Output: key = NULL; value = countryCode, numberOfCustomers, minTransTotal, maxTransTotal **Source Code (Part 1):** Project1/Q3_WritingMapReduceJobs/Q3.2_Query2/CountryCodeGrouping_Part1.java **Output (Part 1):** Project1/Q3_WritingMapReduceJobs/ Q3.2_Query2/FinalOutputQ3bpt1/part-r-00000 **Source Code (Part 2):** Project1/Q3_WritingMapReduceJobs/Q3.2_Query2/CountryCodeGrouping_Part2.java **Output (Part 2):** Project1/Q3_WritingMapReduceJobs/ Q3.2_Query2/FinalOutputQ3bpt2/part-r-00000
Q3.3	Fully Working	*Query Plan:* Customer Map Logic: 4. For a given record, do parsing and extract fields 5. Key = ID value 6. Value = “Customer”, Age value, Gender value Transaction Map Logic: 4. For a given record, do parsing and extract fields 5. Key = CustID value 6. Value = “Transaction”, TransTotal Reduce Logic: 4. Separate array based on dataset source, either “Customer” or “Transaction” 5. Join records

		6. Output: key = NULL; value = ID, Age, Gender, TransTotal Map 2 Logic: 4. For a given record, do parsing and extract fields 5. Key = AgeRange, Gender value 6. Value = TransTotal Reduce 2 Logic: 6. Get MinTransTotal for each AgeRange & Gender group 7. Get MaxTransTotal for each AgeRange & Gender group 8. Get AvgTransTotal for each AgeRange & Gender group 9. Output: key = NULL; value = AgeRange, Gender, MinTransTotal, MaxTransTotal, AvgTransTotal **Source Code (Part 1):** Project1/Q3_WritingMapReduceJobs/Q3.3_Query3/AnalyticsTask_Part1.java **Output (Part 1):** Project1/Q3_WritingMapReduceJobs/ Q3.3_Query3/FinalOutputQ3cpt1/part-r-00000 **Source Code (Part 2):** Project1/Q3_WritingMapReduceJobs/Q3.3_Query3/AnalyticsTask_Part2.java **Output (Part 2):** Project1/Q3_WritingMapReduceJobs/ Q3.3_Query3/FinalOutputQ3cpt2/part-r-00000
Q4.1	Fully Working	*Query Plan:* 1. Load Customers dataset and set fields 2. Load Transactions dataset and set fields 3. Join Customers and Transactions datasets on ID and CustID 4. Group by customer and calculate the number of transactions 5. Find the minimum transaction count by ordering by transaction count and selecting the minimum 6. Check for other customers with the minimum transaction count 7. Print name and transaction count for all customers with the minimum transaction count

		Source Code: Project1/ Q4_WritingApachePigJobs /Q4.1_Query1/PigQuery1.pig **Output:** Project1/ Q4_WritingApachePigJobs /Q4.1_Query1/FinalOutputQ4a/part-m-00000
Q4.2	Fully Working	*Query Plan:* 1. Load Customers dataset and set fields 2. Get only necessary fields: ID and CountryCode 3. Group by CountryCode 4. Count the number of unique customer IDs 5. Select country codes that have greater than 5,000 customers or less than 2,000 customers 6. Print the country codes and their customer numbers **Source Code:** Project1/ Q4_WritingApachePigJobs /Q4.2_Query2/PigQuery2.pig **Output:** Project1/ Q4_WritingApachePigJobs /Q4.2_Query2/FinalOutputQ4b/part-m-00000
Q4.3	Fully Working	*Query Plan:* 1. Load Customers dataset and set fields 2. Load Transactions dataset and set fields 3. Get only necessary fields: id, age, and gender 4. Put age into corresponding group and name this ageGroup 5. Get only necessary fields: CustID, TransTotal 6. Join Customers and Transactions datasets on ID and CustID 7. Group by ageGroup and gender 8. Calculate min trans total, max trans total, and average trans total 9. Print the result **Source Code:** Project1/ Q4_WritingApachePigJobs /Q4.3_Query3/PigQuery3.pig **Output:**

		Project1/ Q4_WritingApachePigJobs /Q4.3_Query3/ FinalOutputQ4b/part-m-00000