**TPC-H Query 5 Report**

**Objective**: To implement TPCH Query 5 on TPCH data using C++ multithreading without utilizing a database management system.

**Approach**: To implement TPCH Query 5 using C++, I used following approaches –

* First, I setup TPC-H v3.0.1 and compile, then generate data using dbgen with scale factor 0.1, 1, 2 … 10.
* Then I create structures for all tables like Customer, Order, LineItem, Supplier, Region and Nation.
* Then I parse all tables data including customer.tbl, orders.tbl, lineitem.tbl, supplier.tbl, region.tbl and nation.tbl and store all data in C++ vector.
* Then I create a function processQuery to perform given query and calculate revenue.
* Then Create a thread manager to execute processQuery parallelly for multiple chunks created according to the given number of threads.
* After that print all the result data and also store in a location provided by user during execution of program.

**Challenges encountered**: During the development, I face following challenges –

* The first challenge was to parse .tbl data efficiently So that I can perform query on it. So I use C++ file system to read data and store in C++ vector.
* In this approach, process was taking too much time to parse data, So I divided data into chunks and use multithreading to perform task parallelly and the execution speed increased by 50%.
* Then I face challenge to process data so that I can execute query efficiently. For this purpose, I create a function processQuery and a function threadManager to manage all threads.
* In thread manager, I divide the process in chunks according to the number of threads given and execute task parallely.

**Lessons learned during the implementation**: I learned following lessons during implementation –

* I learned the use of C++ file system to parse data.
* I learned C++ multithreading and brush up my knowledge of parallel processing.
* I was creating a CLI using bash, during the CLI implementation, it brush up my bash language knowledge like how can take user input in bash, also how to set default values etc.

**Declaration**: This project can perform TPCH Query 5 without using any Database management system. Project uploaded to GitHub: <https://github.com/vikas7754/tpch_query5>

**Future Scope**: Currently it’s taking 25s to query 100MB data, I am thinking to improve the execution speed So that it can perform large data very quickly.

**Note**: Execution guide available in file README.md