FASTcommerce-Business_Intelligence

USAID BIN REHAN 20K-0297
UMER WASI 20K-0318
HUZAIFA JAWAD 20K-0175

NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES - FAST

1. Practical Demonstration of Skills on Data Structures:

Link to the 5+ GB sample dataset used

The HashTable data structures are programmed using a class that has a 2 dimensional data member of user-defined Node class that consists of data members based on the columns, initialized using either default or parameterized constructor. The 2D hashtable data member is allocated and deallocated in the constructor and destructor respectively. The class includes member functions of insertion where the process of hashing is performed, that is the row value of the column considered as a unique key is passed to the hash-function which returns the remainder value after the key value is divided by table size (declared in macro in this case) which is used for the separate chaining. The other member function is the searching where the User_ID or Product_ID to be searched is compared with each key to return true if found.

The queue data structure is imported from the standard template library. Its push, pop and isempty built-in procedures are implemented.

The stack data structure is imported from the standard template library. Its push, pop and isempty built-in functions are demonstrated.

The vector data structure is imported from the standard template library. Its push_back function is used.

2. Data Structures for Handling Large Datasets:

The 2 user-defined HashTable data structures, named HashMap and HashP, utilize a key that is based on the prime number just greater than the number of rows in the dataset. The HashMap is used to process and analyze data using the values in the User_ID column as key value whereas the HashP does the same using the values in the Product_ID column.

The dynamically allocated queue data structure, from the standard template library, provides the functionality of storing the list of products of the user with most interactions, without occupying any unnecessary memory.

The stack data structure, from the standard template library, is used to maintain a list of operations for all the instances of user activities a method wants to store, where the Last-In-First-Out mechanism provides undo or return back to access the previous instance action.

The vector data structure is used with stringstream library's istringstream function to convert string to integer.

These data structures are demonstrated on a big dataset (occupying over 1 GB of memory) comprising over 1 Million rows.

3. Practical Importance of Problem:

Big Data Analysis is a field of processing and analyzing datasets that are too large (occupying multiple gigabytes of hard-disk) or complicated to be dealt with using traditional methods. Big datasets offer greater statistical power for analytics and more training samples to a machine learning model for increased accuracy.

Current usage of the term Big Data refers to the use of analytics, in this case user-behaviour analytics, to extract meaningful insights from big data. The data exists in several formats depending on its structuring as in this case, structured as a comma separated values (CSV) file.

In the age of information overload, companies having excess of big data and lack of resources and optimized methods to process and analyze it, has led to several economic losses.

One of the booming sectors facing this problem is e-commerce that comprises organizations ranging from solopreneurs to large-scale corporations, many of whom have access to their customer behaviour and store it inform of big data. The problem faced by each is the lack of an efficient method to process and analyze it, that is, a method which requires less time and affordable hardware investment, while producing maximum results.

In order to achieve this, a fast low-level turing complete programming language must be used that has functionality available to read the dataset and process it without use of any expensive processors.

4. Domain Preferences:

E-commerce is defined as transactions of products over the Internet. It is a $ 4 billion industry that will only grow in value over time, as it thrives in the majority of global crises and will rise rapidly with Web 3.0 technologies.

The high growth of e-commerce over other sectors is due to a number of reasons. Firstly,

lower capital is required for higher output. Secondly, the ease of global scalability increases with the development of digital applications. Moreover, it provides the ability to collect customer data for customer-centric decision-making. Furthermore, it is an opportunity of 24/7 passive income, by automating salesperson interactions through the use of advancing chatbots that self-improve using the collected data. Lastly, the text-based retargeting, through advertisement posts on social platforms is considered less annoying than its cold calling counterpart.

5. Quality of Solution:

FASTcommerce-Business_Intelligence is a big data analytics console application aimed at fulfilling the need of a fast and low system requirement business intelligence software for e-commerce enterprises who collect their website users' behaviour data for optimizing their website's user experience and monitor the product demand for linear programming of supply.

The program is built using the fastest object-oriented programming language, C++, and its "bits\stdc++.h", "windows.h" and "conio.h" header files as dependencies. It only requires a command prompt console to run the executable file. Since C++ programs are directly compiled without need for any interpretation, it does not need any graphics processing unit.

Its back end comprises the most efficient data structure, Hash Table, whose average time complexity is O(1), that is, it takes constant time to lookup (load, process and search) the data regardless of the size. The worst case time complexity and the time complexity for greatest and least comparisons is O(n), that is, the time taken for lookup depends on the number of rows in the Hash Table.

Despite being a console application, its user interface has been designed as minimalist yet appealing, while the clearly defined actions lead to an optimized user experience, hence lower learning curve and higher productivity.

The main prompt offers 3 instructions:
i. Search with respect to User_ID
ii. Search with respect to Product_ID
iii. Exit

Both of the first 2 instructions leads to a prompt dispensing their respective functions, including but not limited to displaying, the user who bought most products, who bought most and least priced product, the most and the least priced product and the category it

belongs to and the name and category of products with most and least sales. This indefinite list of functions has a potential of increased functionality through open-source contributions.

6. Evaluation and Presentation:

The result of testing each functionality using software testing methods were compile-time and run-time error free. The time complexity is also the desired result.

7. Conclusion:

FASTcommerce-BI is an efficient solution to the problems occurring due to lack of optimized big data processing and analysis aimed at organizations, who want maximum returns on minimum time and hardware investment, who want to scale to a customer-centric culture.