Visvesvaraya Technological University
Belagavi-590 018, Karnataka

A Mini Project Report on

"Implementation of INDEXING for LOCATION Data Set"

Mini Project Report submitted in partial fulfillment of the requirement for the File Structure Lab [17ISL68]

Bachelor of Engineering
In
Information Science and Engineering
Submitted by
PHALGUNI V[1JT17IS025]

Under the guidance of

Mr.Vadiraja.A

Assistant Professor, Department of ISE

Department of Information Science and Engineering
Jyothy Institute of Technology
Tataguni, Bengaluru-560082

CERTIFICATE

Certified that the mini project work entitled **"Implementation of Indexing on Location dataset"** carried out by **Phalguni V [1JT17IS025]** bonfire student of Jyothy Institute of Technology, in partial fulfilment for the award of Bachelor of Engineering in Information Science and Engineering department of the Visvesvaraya Technological University, Belagavi during the year 2020-2021. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

Mr .VadirajaA   **Dr. HarshvardhanTiwari**

Guide, Asst.Professor   Associate. Professor & HOD
Dept.Of ISE   Dept.Of ISE

External Viva Examiner: Signature with Date:

1.

2.

ACKNOWLEDGEMENT

ABSTRACT

Indexing is the process of associating a key with the location of a corresponding data record. An external sort typically uses the concept of a key sort, in which an index file is created whose records consist of key pairs. Here, each key is associated with a pointer to a complete record in the main database file. The index file could be sorted or organised using a tree structure, thereby imposing a logical order on the records without physically rearranging them. Each record of a database normally has a unique identifier, called the primary key. A particular key value might be duplicated in multiple records, is called a secondary key. The secondary key index will associate a secondary key value with the primary key of each record having that secondary key value .The full database might be searched directly for the record with that primary key, or there might be a primary key index that relates each primary key value with a pointer to the actual record on the disk. In this case, the primary index provides the location of the actual record on disk, while the secondary disk indices refer to the primary index. Indexing is an important technique for organising large databases.

Table of Contents

CHAPTER 1
INTRODUCTION

1.1 Introduction to File Structures

File Structures is the organization of data in secondary storage device in such a way that minimize the access time and the storage space. A File Structure is a combination of representations for data in files and of operations for accessing the data. A File Structure allows applications to read, write and modify data. It might also support finding the data that matches some search criteria or reading through the data in some particular order.

In computing, a file system or file system controls how data is stored and retrieved. Without a file system, information placed in a storage medium would be one large body of data with no way to tell where one piece of information stops and the next begins. By separating the data into pieces and giving each piece a name, the information is easily isolated and identified. Taking its name from the way paper-based information systems are named, each group of data is called a "file". The structure and logic rules used to manage the groups of information and their names is called a "file system".

There are many different kinds of file systems. Each one has different structure and logic, properties of speed, flexibility, security, size and more. Some file systems have been designed to be used for specific applications. For example, the ISO 9660 file system is designed specifically for optical discs.

File systems can be used on numerous different types of storage devices that use different kinds of media. The most common storage device in use today is a hard disk drive. Other kinds of media that are used include flash memory, magnetic tapes, and optical discs. In some cases, such as with tmpfs, the computer's main memory (random-access memory, RAM) is used to create a temporary file system for short-term use.

1.2 Indexing

Index : A structure containing a set of entries, each consisting of a key field and a reference field, which is used to locate records in a data file.
Key Field : The part of an index which contains keys.
Reference Field: The part of an index which contains information to locate records.

- An index imposes order on a file without rearranging the file.
- Indexing works by indirection.

A Simple Index for Entry-Sequenced Files
An index in which the entries are a key ordered linear list.

- Simple indexing can be useful when the entire index can be held in memory.
- Changes (additions and deletions) require both the index and the data file to be changed.
- Updates affect the index if the key field is changed, or if the record is moved.

- An update which moves a record can be handled as a deletion followed by an addition.

Indexing to Provide Access by Multiple Keys
A search key other than the primary key.

An index built on a secondary key.

Secondary indexes can be built on any field of the data file, or on combinations of fields.

Secondary indexes will typically have multiple locations for a single key.

Changes to the data may now affect multiple indexes.

The reference field of a secondary index can be a direct reference to the location of the entry in the data file.

The reference field of a secondary index can also be an indirect reference to the location of the entry in the data file, through the primary key.

Indirect secondary key references simplify updating of the file set.

Indirect secondary key references increase access time.

Retrieval Using Combinations of Secondary Keys

The search for records by multiple keys can be done on multiple index, with the combination of index entries defining the records matching the key combination.

If two keys are to be combined, a list of entries from each key index is retrieved.

For an "or" combination of keys, the lists are merged.

I.e., any entry found in either list matches the search.

For an "and" combination of keys, the lists are matched.

I.e., only entries found in both lists match the search.

Improving the Secondary Index Structure: Inverted Lists

An index in which the reference field is the head pointer of a linked list of reference items.

Selective Indexes: An index which contains keys for only part of the records in a data file.

Binding: The association of a symbol with a value. A condition in which items accessed temporally close are also physically close.

1.3 Python

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. □

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object oriented, imperative, functional and procedural, and has a large and comprehensive standard library. □ Python interpreters are available for many operating systems. C Python, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. C Python is managed by the non-profit Python Software Foundation.

CHAPTER 2

IMPLEMENTATION

2.1 Basic operations on Indexing

- Insertion
- Deletion
- Searching by Primary Index
- Searching by Secondary Index
- Printing Indexes

2.2 Search

The structure used in this mini project to store the data is dictionary, hence searching is done in through traversal of elements in the dictionary.Dictionary in Python is an unordered collection of data values, used to store data values like a map, which unlike other Data Types that hold only single value as an element, Dictionary holds key:value pair.

Searching is done using Brute Force method, Iterative loop is used to search through the elements to find the specific element. For this, we just use naive check and compare and return the result once we find the suitable match and break for rest of dictionaries.

Searching is done using both primary key(Pin Code) and secondary key(Place) which achieves both primary indexing and secondary indexing.

2.3 Insertion

Insertion is a simple If-else statement where, if the place does not exist at a given pin code, a new location can be inserted else it displays an error message. The simplicity of the code is to allow the user to insert a new location for any given place for a given location data set of a country.

2.4 Deletion

Deletion is also a simple If-else statement; Deletion of a place is done if it exists in the given data set, else an error message is displayed. This is a quick process that does not require much time as it does not traverse through the dictionary but to the specific index.

CHAPTER 3
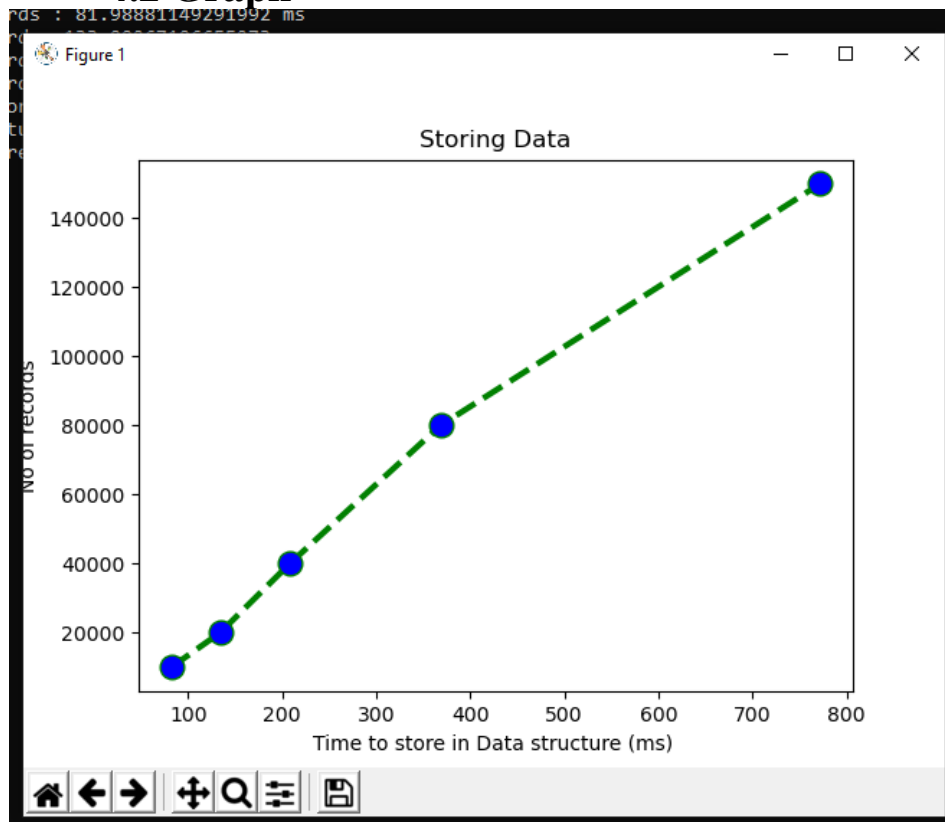# RESULT AND ANALYSIS

4.1 Indexing

In the above image, for the given dataset Location, primary index pincode is depicted in black box showing a 6 digit code for Indian pin code format. The blue box shows the secondary index Places, for the respective pincode.

4.2 Graph

Time analysis for Storing the location dataset in the indexes for all the operations to be performed for the different number of records.

4.3 Indexing time

The image displays to user the time for constructing the indexes, and the various options for options that can be performed on the indexes.

4.4 .1 Insertion(When place doesn't exist)

For Insertion, the pincode is entered and then the place is entered. Time analysis is done for insertion, calculating the time taken. The place 'Isro Layout' is added for pis code 560078.

4.4 .2 Insertion(When place already exists)

```
Your data has been entered
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
4
Enter the pin: 560078
The place is Kumaraswamy Layout
The place is Yelachenahalli
The place is Jp Nagar Iii Phase
The place is JP Nagar
Time taken for the place to be searched :   54.35490608215332 ms
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
1
Enter the pin: 560078
Enter the place name : JP Nagar
This place and pin already exists
Time for the place to be inserted :   0.0 ms
Your data has been entered
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
_
```

If the place, with a specific pincode, the user is trying to insert into the location dataset, the message is displayed the it already existed and replication does not happen.
Here, 'JP Nagar' place with pincode 560078 is being added but it already exists hence is not duplicated.

4.5 Search by Pin

```
This place does not exist
Time taken for the place to be searched :    11.92474365234375 ms
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
4
Enter the pin: 560078
The place is Kumaraswamy Layout
The place is Yelachenahalli
The place is Jp Nagar Iii Phase
The place is J P Nagar
The place is Isro Layout
Time taken for the place to be searched :    73.41790199279785 ms
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
_
```

When the pin code 560078 is searched, all the places with the pincode are displayed
including the place we added in the last step, 'Isro Layout'. The time take is also
calculated.

4.6 Search by Place

```
Time taken for the place to be searched :    ...
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
3
Enter the place name to be searched: Isro Layout
The pin numbers of Isro Layout are as follows
560078
Time taken for the place to be searched :    0.0 ms
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
```

When the place 'Isro Layout' is searched the respective pincode 560078 is displayed.
This implements the secondary indexing. The time analysis for search operation is
also done.

4.7 Deletion

```
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
2
Enter the place name: Isro Layout
Deleted Isro Layout successfully
The pins deleted are as follows:
560078
Time for the place to be deleted :   19.114255905151367 ms
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
```

The place we added 'Isro Layout' is deleted along with its respective pincode. The time taken is also calculated.

```
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
3
Enter the place name to be searched: Isro Layout
This place does not exist
Time taken for the place to be searched :   10.279417037963867 ms
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
```

When the deleted place is searched, error message is displayed. Here, 'Isro Layout' which we deleted is not found in the dataset hence, error message is displayed. Along with the time taken for the whole search process.

4.8 Print Indexes


```
241123   Anangpur      4653
604203   Anangur       4654
605103   Anangur       4655
637304   Anangur       4656
639202   Anangur       4657
813611   Anant   4658
844120   Anant Kamtaul 4659
801102   Anantagiripally      4660
764087   Anantapali    4661
615001   Anantapur     4662
754119   Anantapur     4663
759017   Anantapur     4664
742202   Anantapur     4665
721635   Anantapur     4666
756046   Anantapur (Baleswar)  4667
711301   Anantapur (Howrah)    4668
615001   Anantapur Collectorate    4669
615002   Anantapur Engg College    4670
616355   Anantapuram   4671
731242   Anantapurkanaipur    4672
600101   Anantaram     4673
624302   Anantasagaram 4674
622236   Anantavaram   4675
622301   Anantavaram   4676
623315   Anantavaram   4677
635145   Ananthagiri   4678
633436   Ananthagiri   4679
603186   Ananthagiri   4680
605402   Ananthagiri   4681
608206   Ananthagiri   4682
616107   Ananthaiah Gari Palle  4683
624407   Ananthamadugu 4684
603307   Ananthamangalam     4685
632511   Ananthangal   4686
605524   Ananthapalli  4687
634111   Ananthapalli  4688
624225   Ananthapuram  4689
624344   Ananthapuram  4690
616257   Ananthapuram  4691
632315   Ananthapuram  4692
605201   Ananthapuram  4693
616105   Anantharajup R.S.     4694
```

The indexes with respective pincodes and places is printed. The above image shows this with a parallel dataset with all the collective data.


```
796111   Zohmun        154721
796701   Zohnuai       154722
416511   Zolambe       154723
415702   Zombadi       154724
787060   Zonekareng    154725
781024   Zoo Road      154726
737121   Zoom    154727
263601   Zoor    154728
192303   Zoora   154729
335024   Zorawarpura   154730
796321   Zote 'E'      154731
796181   Zote 'S'      154732
443202   Zotinga       154733
796691   Zotlang       154734
795102   Zoupi   154735
795126   Zouzangtek    154736
391165   Zoz     154737
796017   Zuangtui      154738
403726   Zuarinagar    154739
797005   Zubza   154740
362530   Zudvali       154741
193411   Zugu Khyran   154742
522413   Zulakallu     154743
416220   Zulpewadi     154744
382421   Zundal        154745
364490   Zundala       154746
798620   Zunheboto     154747
209121   Zuniya        154748
591230   Zunzurwad     154749
370001   Zura    154750
193224   Zurhama       154751
425103   Zurkheda      154752
786191   Zutlibari     154753
524152   Zuvvaladinne  154754
523240   Zuvvaleru     154755
523270   Zuvvigunta    154756
Enter your choice
1) Insert new place and pin
2) Delete
3) Search by place
4) Search by pin
5) Print the Indexes
0) Exit
```

Prints till the end of the indexes.

CONCLUSION

We have successfully implemented indexing which helps us in administrating the data used for managing the tasks performed. View tables are used to display all the components at once so that user can see all the components of a particular type at once. One can just select the component and modify and remove the component. We have successfully used various functionalities of Python and created the File structures.

Features:
1. Clean separation of various components to facilitate easy modification and revision.
2. All the data is maintained in a separate file to facilitate easy modification
3. All the data required for different operations is kept in a separate file.
4. Quick and easy saving and loading of database file.

REFERENCES

The information about indexing was gathered by referring to the following sites:

- ☐ Wikipedia(www.wikipedia.org)

- Stackoverflow(stackoverflow.com)

- Michael J.Folk, Bill Zoeclick, Greg Riccardi:File Structures-An Object Oriented Approach with C++,3rd Edition,Pearson Education ,1998.

- Scott   Robert   Ladd:C++   Components   and   Algorithms,BPB Publications,1993