**Big Data Analytics**

**Homework 1 (Introduction to Big Data, B+-Tree, and Extensible Hashing)**

**Name:** Sreenath Reddy Kurukunda

**Student id:** 811211580

1. What are the 3V features of big data? Please briefly describe them. [15 points]  
  
**ANSWER:**  
The 3V features of the Big Data

* Variety
* Volume
* Velocity

Variety: - The variety is defined as the large amount of data is collected from different sources such as organizations and many more. This data can be of any format for example pdf, word, csv, video, images (jpg, png). This variety of data can be in structured, unstructured, and semi structured data. This variety of data is processed using the Big Data.  
Volume: - In this volume the organizations will collects and evaluated and process the data is called as data volume the organizations collect the data in different varieties, so the storage of the data is also increased by each second. For example, amazon is storing every single person data which is around 2000+ petabytes till now and the data is increased very second, to process this data usually need different kinds of processing techniques and technology.   
Velocity: - The velocity is that how quickly the rate at which the data is create and distributed and is coming in, for example the greater number of people sending the request on the internet. The speed of the data which is collected will also be increased and process the data or request and the more it will be retained. To handle the amount of data in less time is called Velocity of data.

2. In the following figure, which is sparse index? Which is dense index? [20 points]

**ANSWER:**Sparse index: -In the above index A is the Sparse index. Sparse Index is that which has only some index of the key values in the database file. The sparse index can only be built on the ordered filed of the database file.  
Dense index: - In the above index B is the Dense index. Dense index is that which has all the key values of the given database file. The Dense index can be built for all ordered and unordered field of database file.   
  
  
3. Please tell the differences between B-tree and B+-tree. (**Hint:** please use Google to find the structure of B-tree index) [20 points]  
  
**ANSWER:**

**Index A:**

**Index B:**



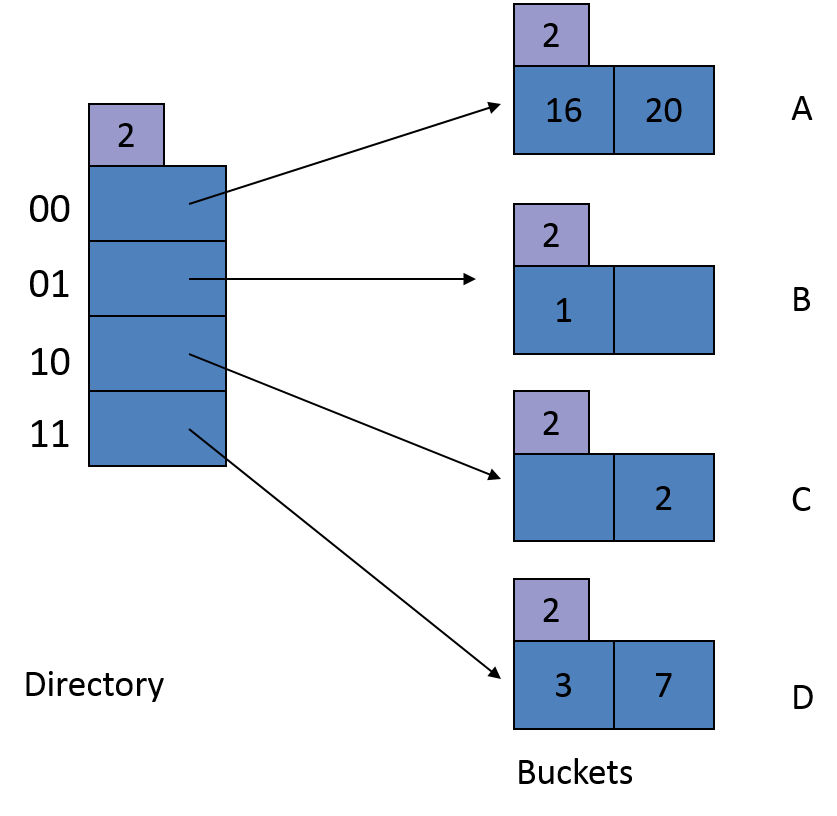
B -tree:  
🡪 Duplicates keys are not present in the B -tree.

🡪 The B -tree will take more memory and computation.  
🡪 The data in the leaf nodes is stored in a linked list.  
🡪 There is no topic such as redundant keys.  
🡪 Does not keep a copy of any keys.  
🡪 Each node is associated with data pointer.  
  
B+-tree

🡪Duplicates keys are maintained, and all the nodes are present at leaf. As leaf nodes have data pointer, we can easily traverse the tree.

🡪 Will occupy very less memory.  
🡪 The data in the leaf nodes is not stored in a linked list.  
🡪 There are redundant keys for some cases.  
🡪 Contains duplicate of keys.  
🡪 data pointers while the root nodes are not data pointers.

4. In the extensible hash index (each bucket can hold at most 2 items) below, if we want to insert 19, what will be the structure of the hash index after the insertion? [20 points]



*h*(1) = 000**01**

*h*(2) = 000**10**

*h*(3) = 000**11**

*h*(7) = 001**11**

*h*(16) = 100**00**

*h*(18) = 100**10**

*h*(19) = 100**11**

*h*(20) = 101**00**

**ANSWER:**The hash index can hold only at most 2 buckets values. If we want to add or insert a hash value h(19), this only possible to insert with h(3) and h(7). But the h(3) and h(7) has the two right most bits in h(19) that is, 11. The value h(19) will be added to the bucket D but the bucket D is already filled with hash values h(3) and h(7) which are also ending with 11.it is not possible to add more than 2 values in a bucket then the last three digits will be considered values of h(3), h(7) and h(19) which are 011,111,011.  
the values of h(3) and h(19) are both similar 011. Which will be filled in bucket D. whereas the h(7) will move and filled in the next bucket E.

2

|  |
| --- |
| 3 |

3 19

2

D

CDDD

B

2

16 20

AB

2

2

1

|  |
| --- |
| 000 |
| 001 |
| 010 |
| 100 |
| 110 |
| 101 |
| 011 |
| 111 |

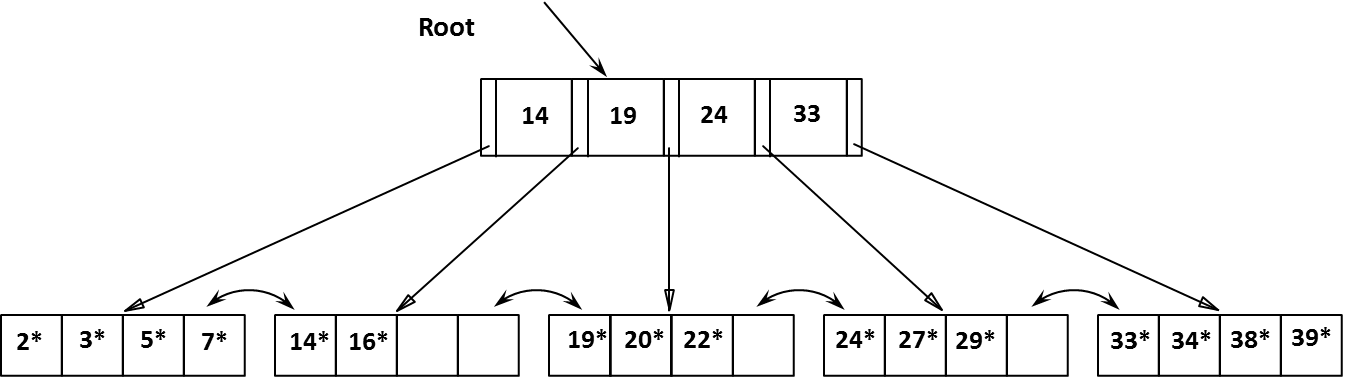
2

5. Assuming that the node capacity of B+-trees is 4, please draw the updated B+-trees upon insertions. [25 points]

7

E

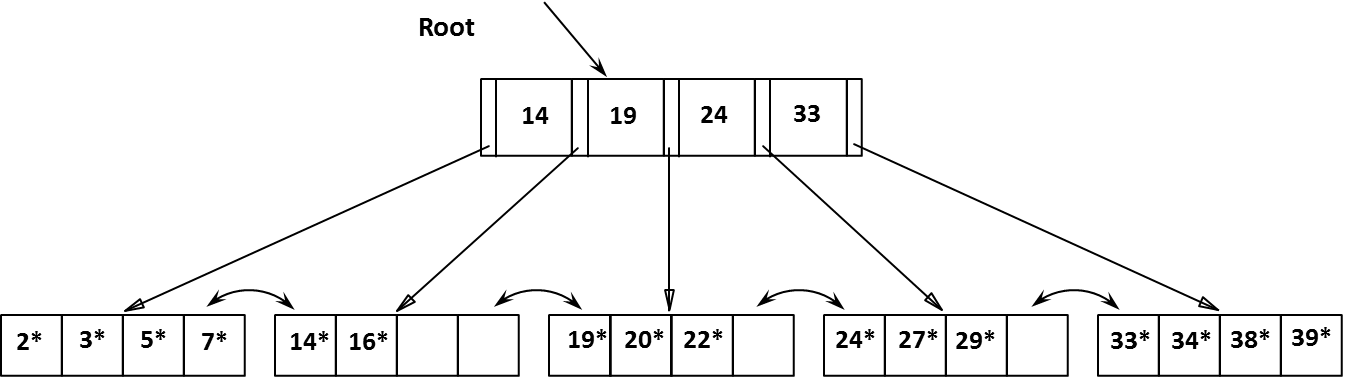
5(a). Insert tuples with key 23\* into the following B+-tree. [10 points]

  
  
**ANSWER:**

A piece of paper with writing

Description automatically generated with medium confidence

5(b). Insert tuples with key 8\* into the following B+-tree. [15 points]



## ANSWER:Text, letter Description automatically generated