COMP9315 ASST2 REPORT

Comparison of SIMC Indexing Methods

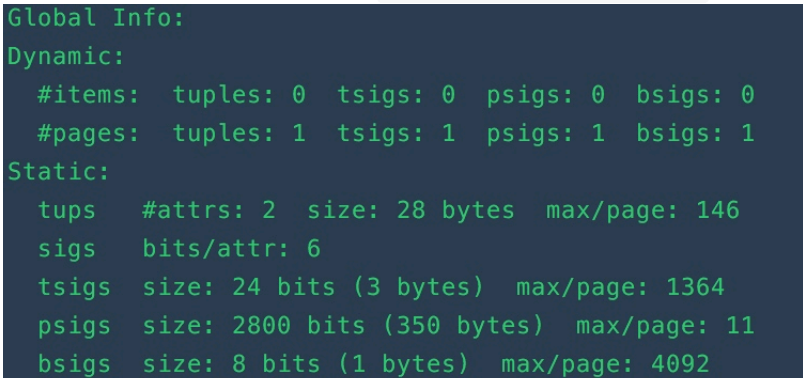
This report is focusing on comparing effectiveness about three mainly SIMC indexing methods. As known to all, for querying a particular record (AKA tuple), no matter what type it is, the dummy way is to look through all the records in the pages in the database management. However, the number of records and pages can be relative large that costs a harsh situation on reading. Hence, SIMC's reducing pages read as well as tuple read become significant.

In the next few paragraphs, this report is going to demonstrate how many tuples and pages are read in different scenarios (i.e., different size of tuples and of all kinds of signatures) in order to show the effectiveness of each.

Small Size of All

> Minimum and default values:
> The number of tuples must not be less than 10.
> The number of attributes must be between 1 and 10 (both excluded).
> The probability of false match must be less than 0.1.

Running command with `./create R 10 2 100` comes with



```
Global Info:
Dynamic:
  #items:  tuples: 0  tsigs: 0  psigs: 0  bsigs: 0
  #pages:  tuples: 1  tsigs: 1  psigs: 1  bsigs: 1
Static:
  tups   #attrs: 2  size: 28 bytes  max/page: 146
  sigs   bits/attr: 6
  tsigs  size: 24 bits (3 bytes)  max/page: 1364
  psigs  size: 2800 bits (350 bytes)  max/page: 11
  bsigs  size: 8 bits (1 bytes)  max/page: 4092
```

Initially test with unrepeated data with `./gendata 10 2 | ./insert R`:

1. Open query:

    - Without indexing: The query engine takes all tuples into account and examine all of them.
    - With tuple signature: The tuple signature file shows that only first page has the matching tuple, and since there is only one page after all, the query engine reads not

  only 10 tuples but also 10 tuple signatures.
  - With page signature: There is only one data page in this relation as well as one page. Hence, the query engine reads one page signature which locates to first page of data file. In other words, query read all tuples like above operations do.
  - With bit-slice page signature: THE SAME.

2. ~~Query with one solution~~

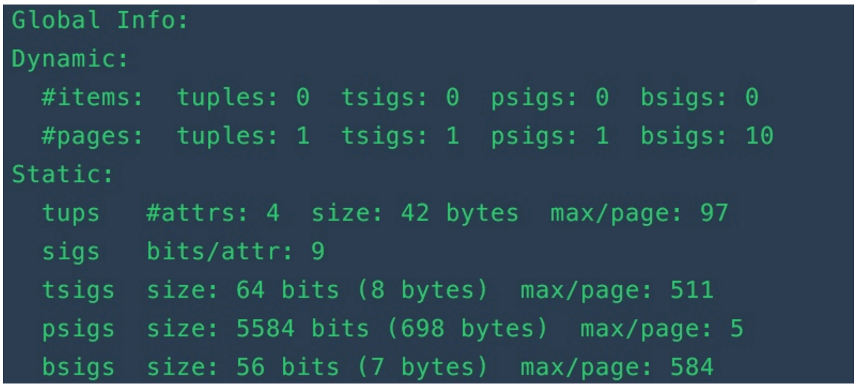
3. ~~Query with many solutions~~

4. ~~Query with multiple values: since there are only two attributes~~

By summarising above, although such a small relation (i.e., tuples will only be hold in one page) is not practical, for those small relations, reading tuple directly from database cost less than using any indexing.

Normal Size of All

Running command with `./create R 5000 4 1000` comes with



```
Global Info:
Dynamic:
  #items:  tuples: 0  tsigs: 0  psigs: 0  bsigs: 0
  #pages:  tuples: 1  tsigs: 1  psigs: 1  bsigs: 10
Static:
  tups   #attrs: 4  size: 42 bytes  max/page: 97
  sigs   bits/attr: 9
  tsigs  size: 64 bits (8 bytes)  max/page: 511
  psigs  size: 5584 bits (698 bytes)  max/page: 5
  bsigs  size: 56 bits (7 bytes)  max/page: 584
```

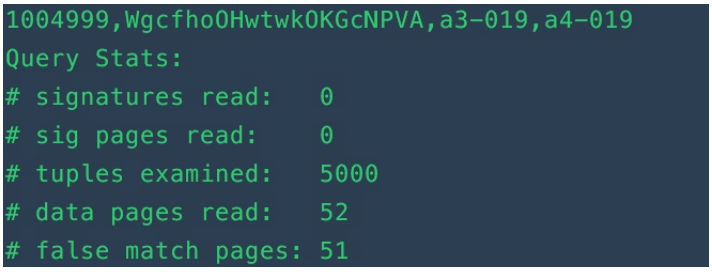
Initially test with unrepeated data with `./gendata 5000 4 | ./insert R`:

1. Open query:

  - Without indexing: The query engine takes all tuples into account and examine all of them.
  - With tuple signature: The tuple signature file shows that all pages have the matching tuple, and there are 52 pages in total, the query engine reads whole 52 pages.
  - With page signature: THE SAME.
  - With bit-slice page signature: THE SAME.
  - In brief, for open queries, the query engine would return all the tuples finally, after all. Query without indexing somehow cost least.

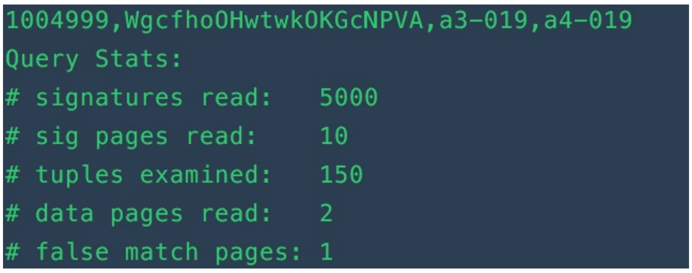
2. Query with one solution: `./select R 1004999,?,?,? [idx]`

- Without indexing: To query the last tuple in the whole database without indexing require query engine obviously read all of tuples in all of pages to find it.

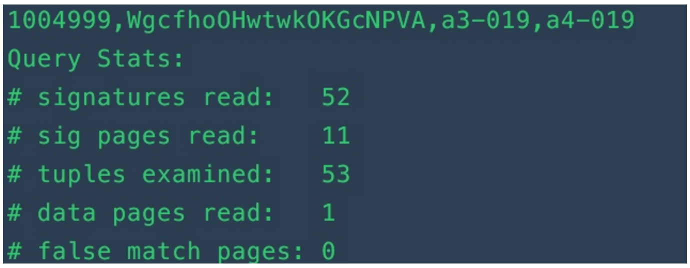
  ```
  1004999,WgcfhoOHwtwkOKGcNPVA,a3-019,a4-019
  Query Stats:
  # signatures read:   0
  # sig pages read:    0
  # tuples examined:   5000
  # data pages read:   52
  # false match pages: 51
  ```

- With tuple signature: With indexing, query becomes a lot easier. However, tuple signature still requires query engine to take all of the tuple signature into account to locate matching pages. Not mention there might be false matches. But it improves the process querying into actual data file because signature tells engine that there is only few possible pages that query tuple might be in.

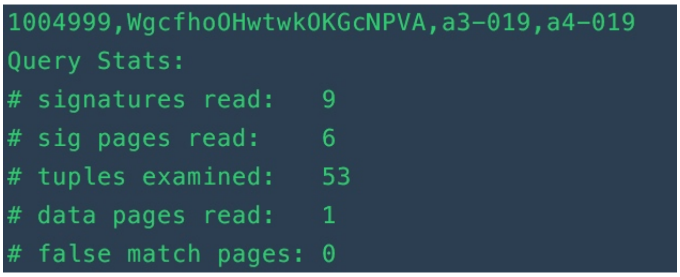
  ```
  1004999,WgcfhoOHwtwkOKGcNPVA,a3-019,a4-019
  Query Stats:
  # signatures read:   5000
  # sig pages read:    10
  # tuples examined:   150
  # data pages read:   2
  # false match pages: 1
  ```

- With page signature: Compared to last option, this improves query process much more. Since, there is only 52 tuple pages leading to 52 page signatures, these 52 page signatures are able to directly tells which pages the query tuple might be in. And finally, query engine goes to those matching pages to look through each tuple in that page to find the genuine matches.

  ```
  1004999,WgcfhoOHwtwkOKGcNPVA,a3-019,a4-019
  Query Stats:
  # signatures read:   52
  # sig pages read:    11
  # tuples examined:   53
  # data pages read:   1
  # false match pages: 0
  ```

- With bit-slice page signature: Compared to page signature, this improves query process into another level. Since possible matching pages now become to the item in the pages. It is no longer necessary to read of the page signatures; yet, all we have to read is only the set bit in Query signature. Not only the signatures read drops a lot, but also signature pages read decreases a few.

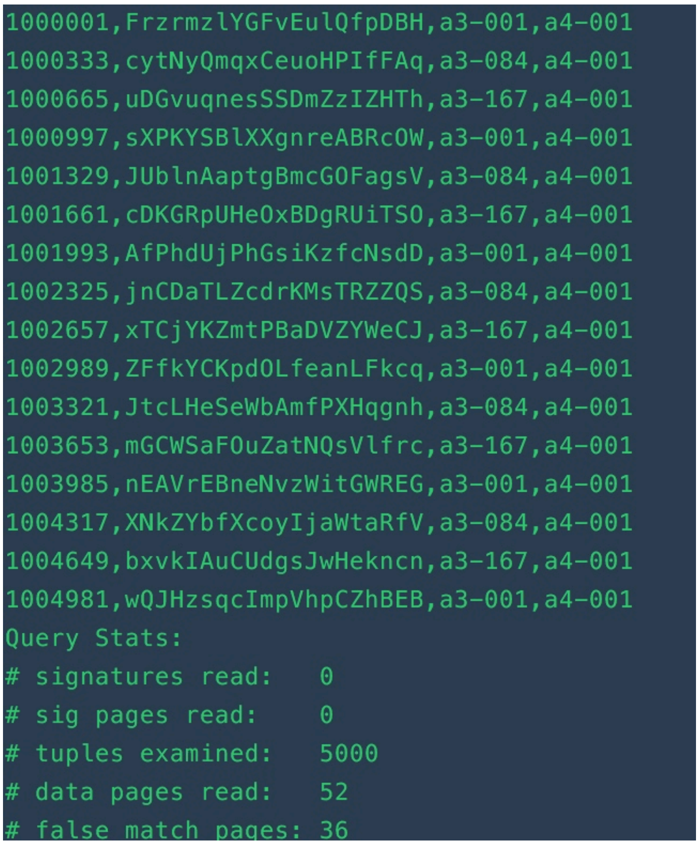


```
1004999,WgcfhoOHwtwkOKGcNPVA,a3-019,a4-019
Query Stats:
# signatures read:   9
# sig pages read:    6
# tuples examined:   53
# data pages read:   1
# false match pages: 0
```

- In brief, for query with one solution, query without indexing cannot be a option with indexing is available. Bit-slice and page signature are definitely prior opportunities to be seized.

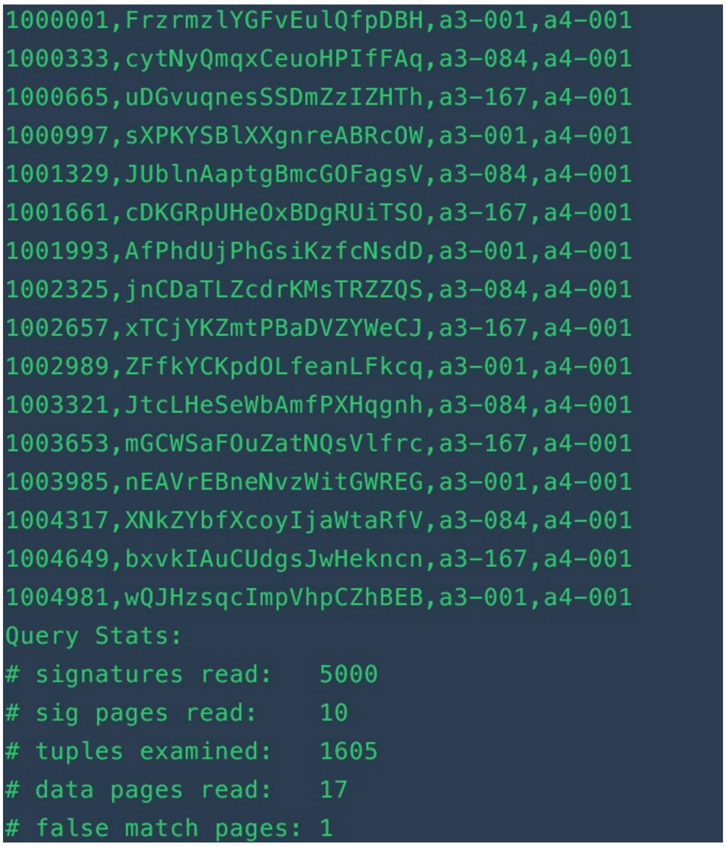
3. Query with many solutions: `./select R "?,?,?,a4-001"`

    - Without indexing:



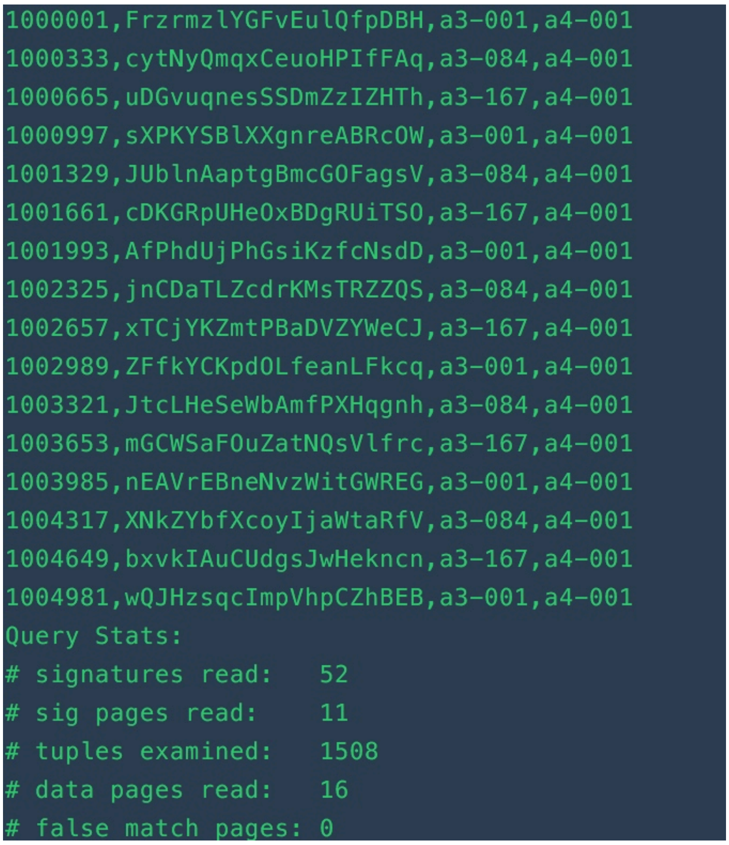
```
1000001,FrzrmzlYGFvEulQfpDBH,a3-001,a4-001
1000333,cytNyQmqxCeuoHPIfFAq,a3-084,a4-001
1000665,uDGvuqnesSSDmZzIZHTh,a3-167,a4-001
1000997,sXPKYSBlXXgnreABRcOW,a3-001,a4-001
1001329,JUblnAaptgBmcGOFagsV,a3-084,a4-001
1001661,cDKGRpUHeOxBDgRUiTSO,a3-167,a4-001
1001993,AfPhdUjPhGsiKzfcNsdD,a3-001,a4-001
1002325,jnCDaTLZcdrKMsTRZZQS,a3-084,a4-001
1002657,xTCjYKZmtPBaDVZYWeCJ,a3-167,a4-001
1002989,ZFfkYCKpdOLfeanLFkcq,a3-001,a4-001
1003321,JtcLHeSeWbAmfPXHqgnh,a3-084,a4-001
1003653,mGCWSaFOuZatNQsVlfrc,a3-167,a4-001
1003985,nEAVrEBneNvzWitGWREG,a3-001,a4-001
1004317,XNkZYbfXcoyIjaWtaRfV,a3-084,a4-001
1004649,bxvkIAuCUdgsJwHekncn,a3-167,a4-001
1004981,wQJHzsqcImpVhpCZhBEB,a3-001,a4-001
Query Stats:
# signatures read:   0
# sig pages read:    0
# tuples examined:   5000
# data pages read:   52
# false match pages: 36
```

    - With tuple signature:



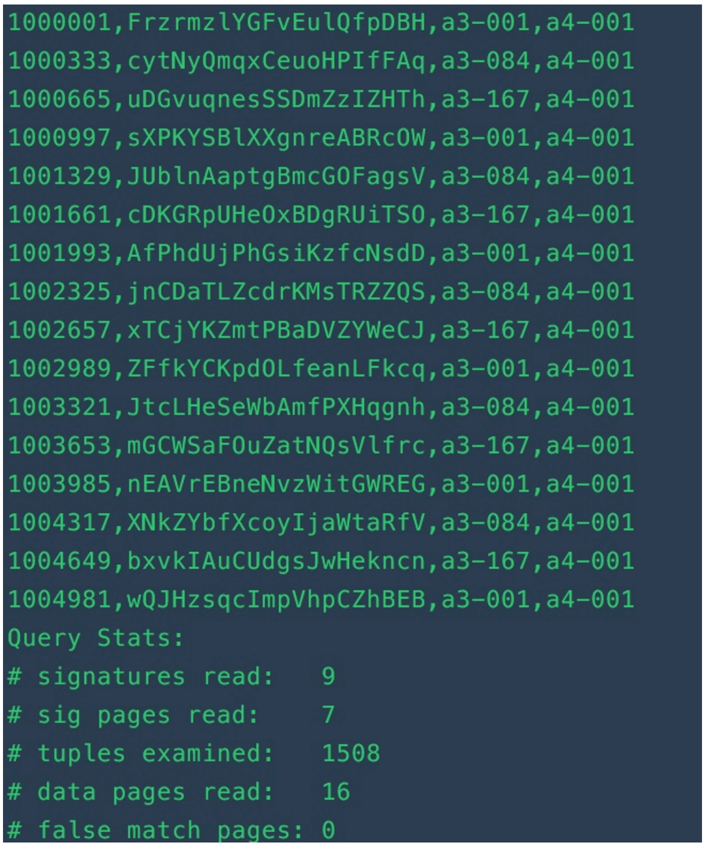
```
1000001,FrzrmzlYGFvEulQfpDBH,a3-001,a4-001
1000333,cytNyQmqxCeuoHPIfFAq,a3-084,a4-001
1000665,uDGvuqnesSSDmZzIZHTh,a3-167,a4-001
1000997,sXPKYSBlXXgnreABRcOW,a3-001,a4-001
1001329,JUblnAaptgBmcGOFagsV,a3-084,a4-001
1001661,cDKGRpUHeOxBDgRUiTSO,a3-167,a4-001
1001993,AfPhdUjPhGsiKzfcNsdD,a3-001,a4-001
1002325,jnCDaTLZcdrKMsTRZZQS,a3-084,a4-001
1002657,xTCjYKZmtPBaDVZYWeCJ,a3-167,a4-001
1002989,ZFfkYCKpdOLfeanLFkcq,a3-001,a4-001
1003321,JtcLHeSeWbAmfPXHqgnh,a3-084,a4-001
1003653,mGCWSaFOuZatNQsVlfrc,a3-167,a4-001
1003985,nEAVrEBneNvzWitGWREG,a3-001,a4-001
1004317,XNkZYbfXcoyIjaWtaRfV,a3-084,a4-001
1004649,bxvkIAuCUdgsJwHekncn,a3-167,a4-001
1004981,wQJHzsqcImpVhpCZhBEB,a3-001,a4-001
Query Stats:
# signatures read:   5000
# sig pages read:    10
# tuples examined:   1605
# data pages read:   17
# false match pages: 1
```

- With page signature:



```
1000001,FrzrmzlYGFvEulQfpDBH,a3-001,a4-001
1000333,cytNyQmqxCeuoHPIfFAq,a3-084,a4-001
1000665,uDGvuqnesSSDmZzIZHTh,a3-167,a4-001
1000997,sXPKYSBlXXgnreABRcOW,a3-001,a4-001
1001329,JUblnAaptgBmcGOFagsV,a3-084,a4-001
1001661,cDKGRpUHeOxBDgRUiTSO,a3-167,a4-001
1001993,AfPhdUjPhGsiKzfcNsdD,a3-001,a4-001
1002325,jnCDaTLZcdrKMsTRZZQS,a3-084,a4-001
1002657,xTCjYKZmtPBaDVZYWeCJ,a3-167,a4-001
1002989,ZFfkYCKpdOLfeanLFkcq,a3-001,a4-001
1003321,JtcLHeSeWbAmfPXHqgnh,a3-084,a4-001
1003653,mGCWSaFOuZatNQsVlfrc,a3-167,a4-001
1003985,nEAVrEBneNvzWitGWREG,a3-001,a4-001
1004317,XNkZYbfXcoyIjaWtaRfV,a3-084,a4-001
1004649,bxvkIAuCUdgsJwHekncn,a3-167,a4-001
1004981,wQJHzsqcImpVhpCZhBEB,a3-001,a4-001
Query Stats:
# signatures read:   52
# sig pages read:    11
# tuples examined:   1508
# data pages read:   16
# false match pages: 0
```

- With bit-slice page signature:



```
1000001,FrzrmzlYGFvEulQfpDBH,a3-001,a4-001
1000333,cytNyQmqxCeuoHPIfFAq,a3-084,a4-001
1000665,uDGvuqnesSSDmZzIZHTh,a3-167,a4-001
1000997,sXPKYSBlXXgnreABRcOW,a3-001,a4-001
1001329,JUblnAaptgBmcGOFagsV,a3-084,a4-001
1001661,cDKGRpUHeOxBDgRUiTSO,a3-167,a4-001
1001993,AfPhdUjPhGsiKzfcNsdD,a3-001,a4-001
1002325,jnCDaTLZcdrKMsTRZZQS,a3-084,a4-001
1002657,xTCjYKZmtPBaDVZYWeCJ,a3-167,a4-001
1002989,ZFfkYCKpdOLfeanLFkcq,a3-001,a4-001
1003321,JtcLHeSeWbAmfPXHqgnh,a3-084,a4-001
1003653,mGCWSaFOuZatNQsVlfrc,a3-167,a4-001
1003985,nEAVrEBneNvzWitGWREG,a3-001,a4-001
1004317,XNkZYbfXcoyIjaWtaRfV,a3-084,a4-001
1004649,bxvkIAuCUdgsJwHekncn,a3-167,a4-001
1004981,wQJHzsqcImpVhpCZhBEB,a3-001,a4-001
Query Stats:
# signatures read:   9
# sig pages read:    7
# tuples examined:   1508
# data pages read:   16
# false match pages: 0
```

- In brief, different from query with one solution, query more than one solutions drops the precision of hash function, though bits is large enough. False matching pages starts to show up.

By summarising above, for a large and normal database, indexing is a necessary for query that reduces reading cost.