Carnegie Mellon University

Lecture #11

ADVANCED DATABASE SYSTEMS

System Catalogs and Database Compression

@Andy_Pavlo // 15-721 // Spring 2018

DATABASE TALK

Oracle In-Memory Database Engine

→ Monday February 26th @ 4:30pm
→ GHC 4401

http://db.cs.cmu.edu/events/db-seminar-spring-2018-ajit-mylavarapu-oracle/

TODAY'S AGENDA

System Catalogs

Compression Background
Naïve Compression
OLAP Columnar Compression

SYSTEM CATALOGS

Almost every DBMS stores their a database's catalog in itself.

→ Wrap object abstraction around tuples.

→ Specialized code for "bootstrapping" catalog tables.

The entire DBMS should be aware of transactions in order to automatically provide ACID guarantees for DDL commands and concurrent txns.

SCHEMA CHANGES

ADD COLUMN:

→ **NSM**: Copy tuples into new region in memory.
→ **DSM**: Just create the new column segment

DROP COLUMN:

→ **NSM #1**: Copy tuples into new region of memory.
→ **NSM #2**: Mark column as "deprecated", clean up later.
→ **DSM**: Just drop the column and free memory.

CHANGE COLUMN:

→ Check whether the conversion is allowed to happen. Depends on default values.

INDEXES

CREATE INDEX:

→ Scan the entire table and populate the index.
→ Have to record changes made by txns that modified the table while another txn was building the index.
→ When the scan completes, lock the table and resolve changes that were missed after the scan started.

DROP INDEX:

→ Just drop the index logically from the catalog.
→ It only becomes "invisible" when the txn that dropped it commits. All existing txns will still have to update it.

SEQUENCES

Typically stored in the catalog. Used for maintaining a global counter

→ Also called "auto-increment" or "serial" keys

Sequences are not maintained with the same isolation protection as regular catalog entries.

→ Rolling back a txn that incremented a sequence does not rollback the change to that sequence.

OBSERVATION

I/O is the main bottleneck if the DBMS has to fetch data from disk.

In-memory DBMSs are more complicated

→ Compressing the database reduces DRAM requirements and processing.

Key trade-off is <u>**speed**</u> vs. <u>**compression ratio**</u>

→ In-memory DBMSs (always?) choose speed.

REAL-WORLD DATA CHARACTERISTICS

Data sets tend to have highly **skewed** distributions for attribute values.

→ Example: Zipfian distribution of the Brown Corpus

Data sets tend to have high **correlation** between attributes of the same tuple.

→ Example: Zip Code to City, Order Date to Ship Date

DATABASE COMPRESSION

Goal #1: Must produce fixed-length values.

Goal #2: Allow the DBMS to postpone decompression as long as possible during query execution.

Goal #3: Must be a **lossless** scheme.

LOSSLESS VS. LOSSY COMPRESSION

When a DBMS uses compression, it is always **lossless** because people don't like losing data.

Any kind of **lossy** compression is has to be performed at the application level.

Some new DBMSs support approximate queries
→ Example: BlinkDB, SnappyData, XDB, Oracle (2017)

ZONE MAPS

Pre-computed aggregates for blocks of data.

DBMS can check the zone map first to decide whether it wants to access the block.

```
SELECT * FROM table
 WHERE val > 600
```

Original Data

val
100
200
300
400
400

Zone Map

type	val
MIN	100
MAX	400
AVG	280
SUM	1400
COUNT	5

COMPRESSION GRANULARITY

Choice #1: Block-level
→ Compress a block of tuples for the same table.

Choice #2: Tuple-level
→ Compress the contents of the entire tuple (NSM-only).

Choice #3: Attribute-level
→ Compress a single attribute value within one tuple.
→ Can target multiple attributes for the same tuple.

Choice #4: Column-level
→ Compress multiple values for one or more attributes stored for multiple tuples (DSM-only).

NAÏVE COMPRESSION

Compress data using a general purpose algorithm. Scope of compression is only based on the data provided as input.

→ LZO (1996), LZ4 (2011), Snappy (2011), Zstd (2015)

Considerations

→ Computational overhead

→ Compress vs. decompress speed.

NAÏVE COMPRESSION

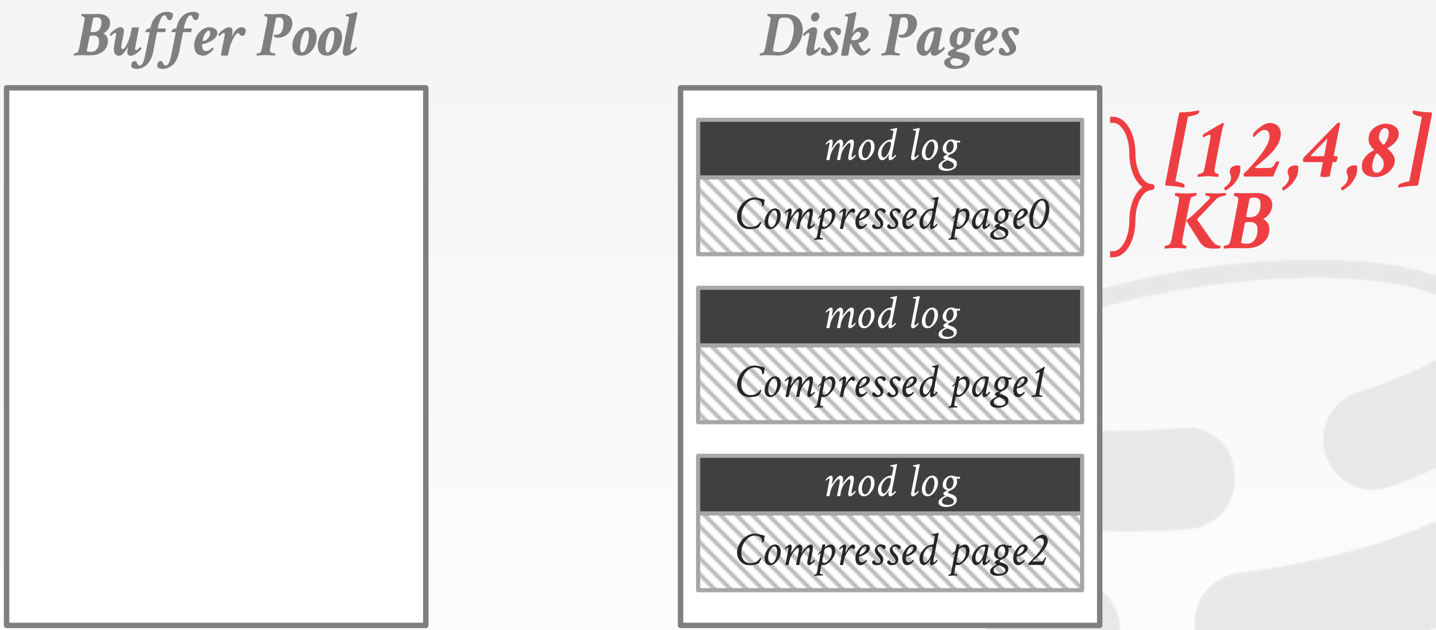
Choice #1: Entropy Encoding

→ More common sequences use less bits to encode, less common sequences use more bits to encode.

Choice #2: Dictionary Encoding

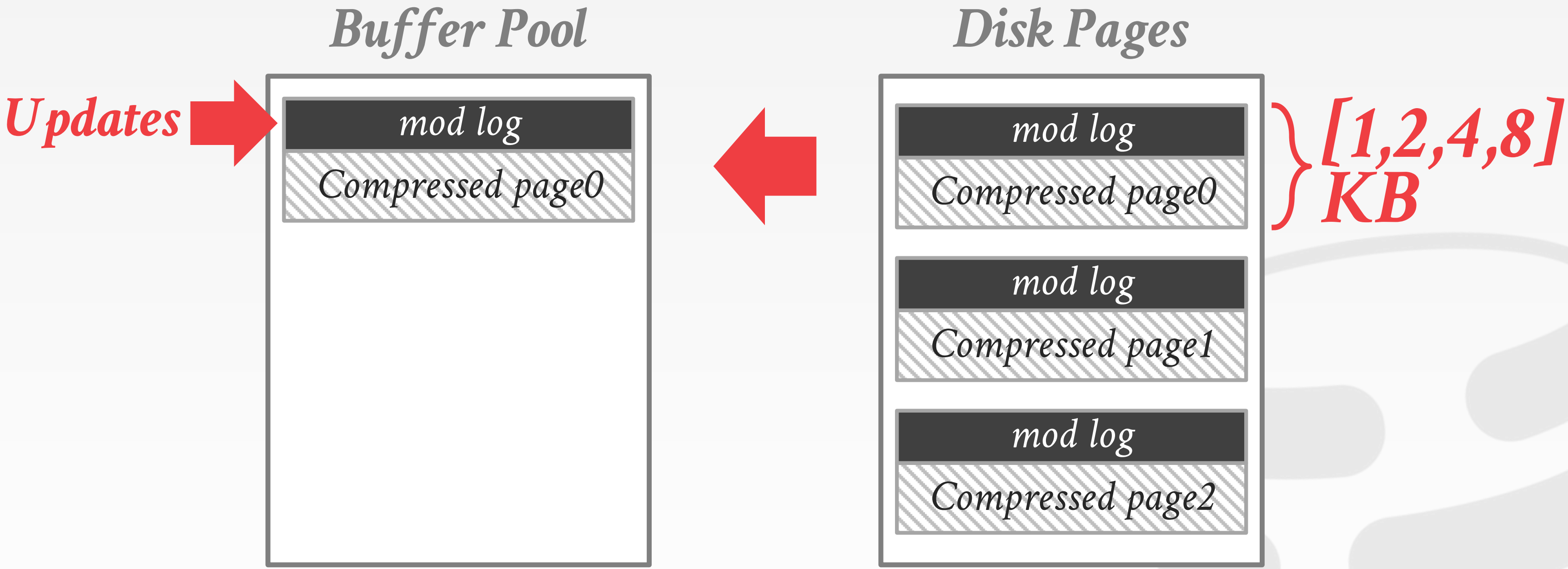
→ Build a data structure that maps data segments to an identifier. Replace those segments in the original data with a reference to the segments position in the dictionary data structure.

MYSQL INNODB COMPRESSION



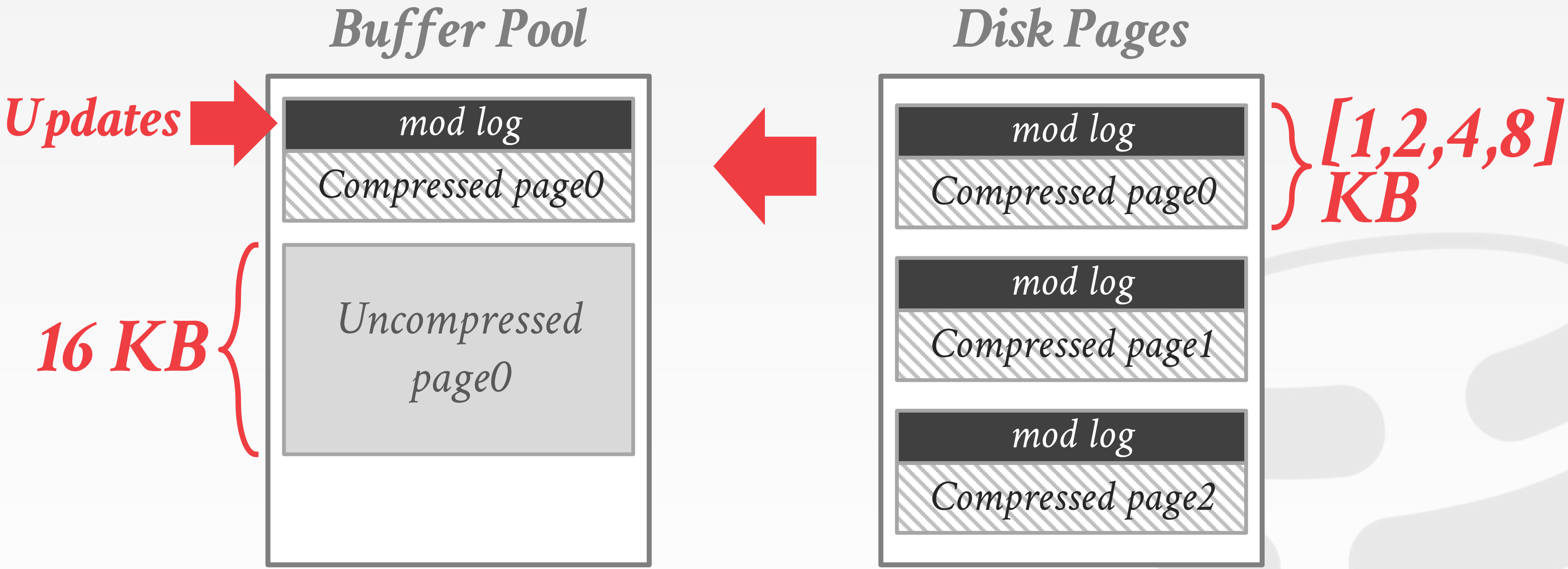
Source: MySQL 5.7 Documentation

MYSQL INNODB COMPRESSION



Source: MySQL 5.7 Documentation

MYSQL INNODB COMPRESSION



Source: MySQL 5.7 Documentation

NAÏVE COMPRESSION

The data has to be decompressed first before it can be read and (potentially) modified.
→ This limits the "scope" of the compression scheme.

These schemes also do not consider the high-level meaning or semantics of the data.

OBSERVATION

We can perform exact-match comparisons and natural joins on compressed data if predicates and data are compressed the same way.

→ Range predicates are more tricky…

```
SELECT * FROM users
 WHERE name = 'Andy'
```

NAME	SALARY
Andy	99999
Prashanth	88888

OBSERVATION

We can perform exact-match comparisons and natural joins on compressed data if predicates and data are compressed the same way.

→ Range predicates are more tricky…

```
SELECT * FROM users
 WHERE name = 'Andy'
```

```
SELECT * FROM users
 WHERE name = XX
```

NAME	*SALARY*
Andy	*99999*
Prashanth	*88888*

NAME	*SALARY*
XX	AA
YY	BB

COLUMNAR COMPRESSION

Run-length Encoding
Bitmap Encoding
Delta Encoding
Incremental Encoding
Mostly Encoding
Dictionary Encoding

RUN-LENGTH ENCODING

Compress runs of the same value in a single column into triplets:

→ The value of the attribute.
→ The start position in the column segment.
→ The # of elements in the run.

Requires the columns to be sorted intelligently to maximize compression opportunities.

DATABASE COMPRESSION
SIGMOD RECORD 1993

RUN-LENGTH ENCODING

Original Data

id	sex
1	M
2	M
3	M
4	F
6	M
7	F
8	M
9	M

RUN-LENGTH ENCODING

Original Data

id	sex
1	M
2	M
3	M
4	F
6	M
7	F
8	M
9	M

Compressed Data

id	sex
1	(M,0,3)
2	(F,3,1)
3	(M,4,1)
4	(F,5,1)
6	(M,6,2)
7	
8	
9	

RLE Triplet
- Value
- Offset
- Length

RUN-LENGTH ENCODING

Original Data

id	sex
1	M
2	M
3	M
4	F
6	M
7	F
8	M
9	M

Compressed Data

id	sex
1	(M,0,3)
2	(F,3,1)
3	(M,4,1)
4	(F,5,1)
6	(M,6,2)
7	
8	
9	

RLE Triplet
- ***Value***
- ***Offset***
- ***Length***

RUN-LENGTH ENCODING

Sorted Data

id	sex
1	M
2	M
3	M
6	M
8	M
9	M
4	F
7	F

Compressed Data

id	sex
1	(M,0,6)
2	(F,7,2)
3	
6	
7	
9	
4	
7	

RLE Triplet
- *Value*
- *Offset*
- *Length*

BITMAP ENCODING

Store a separate Bitmap for each unique value for a particular attribute where an offset in the vector corresponds to a tuple.

→ The i^{th} position in the Bitmap corresponds to the i^{th} tuple in the table.

→ Typically segmented into chunks to avoid allocating large blocks of contiguous memory.

Only practical if the value cardinality is low.

MODEL 204 ARCHITECTURE AND PERFORMANCE
High Performance Transaction Systems 1987

BITMAP ENCODING

Original Data

id	sex
1	M
2	M
3	M
4	F
6	M
7	F
8	M
9	M

BITMAP ENCODING

Original Data

id	sex
1	M
2	M
3	M
4	F
6	M
7	F
8	M
9	M

Compressed Data

id	sex	
	M	F
1	1	0
2	1	0
3	1	0
4	0	1
6	1	0
7	0	1
8	1	0
9	1	0

BITMAP ENCODING

Original Data

id	sex
1	M
2	M
3	M
4	F
6	M
7	F
8	M
9	M

Compressed Data

id	sex: M	sex: F
1	1	0
2	1	0
3	1	0
4	0	1
6	1	0
7	0	1
8	1	0
9	1	0

BITMAP ENCODING: EXAMPLE

```
CREATE TABLE customer_dim (
  id INT PRIMARY KEY,
  name VARCHAR(32),
  email VARCHAR(64),
  address VARCHAR(64),
  zip_code INT
);
```

Assume we have 10 million tuples.
43,000 zip codes in the US.
→ 10000000 × 32-bits = **40 MB**
→ 10000000 × 43000 = **53.75 GB**

Every time a txn inserts a new tuple, we have to extend 43,000 different bitmaps.

DELTA ENCODING

Recording the difference between values that follow each other in the same column.

→ The base value can be stored in-line or in a separate look-up table.

→ Can be combined with RLE to get even better compression ratios.

Original Data

time	temp
12:00	99.5
12:01	99.4
12:02	99.5
12:03	99.6
12:04	99.4

DELTA ENCODING

Recording the difference between values that follow each other in the same column.

→ The base value can be stored in-line or in a separate look-up table.

→ Can be combined with RLE to get even better compression ratios.

Original Data

time	temp
12:00	99.5
12:01	99.4
12:02	99.5
12:03	99.6
12:04	99.4

Compressed Data

time	temp
12:00	99.5
+1	-0.1
+1	+0.1
+1	+0.1
+1	-0.2

DELTA ENCODING

Recording the difference between values that follow each other in the same column.

→ The base value can be stored in-line or in a separate look-up table.

→ Can be combined with RLE to get even better compression ratios.

Original Data

time	temp
12:00	99.5
12:01	99.4
12:02	99.5
12:03	99.6
12:04	99.4

Compressed Data

time	temp
12:00	99.5
+1	-0.1
+1	+0.1
+1	+0.1
+1	-0.2

Compressed Data

time	temp
12:00	99.5
(+1,4)	-0.1
	+0.1
	+0.1
	-0.2

INCREMENTAL ENCODING

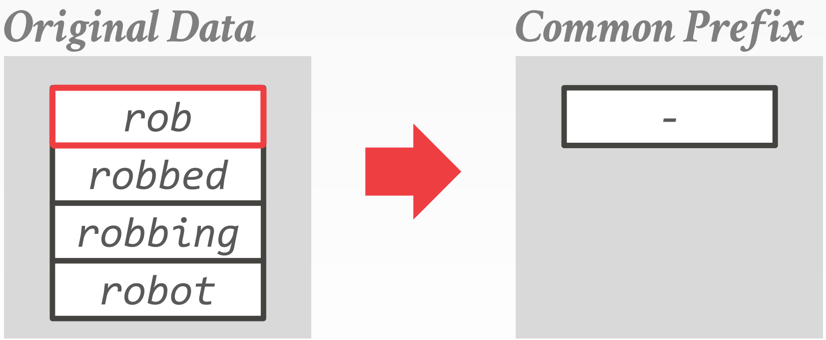
Type of delta encoding whereby common prefixes or suffixes and their lengths are recorded so that they need not be duplicated.
This works best with sorted data.

Original Data

rob
robbed
robbing
robot

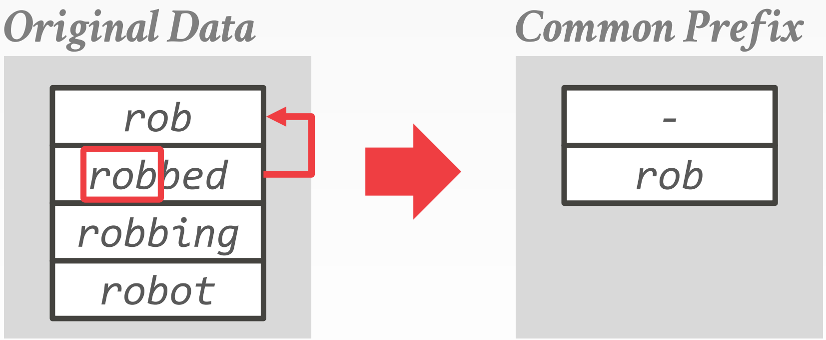
INCREMENTAL ENCODING

Type of delta encoding whereby common prefixes or suffixes and their lengths are recorded so that they need not be duplicated.
This works best with sorted data.



INCREMENTAL ENCODING

Type of delta encoding whereby common prefixes or suffixes and their lengths are recorded so that they need not be duplicated.
This works best with sorted data.



INCREMENTAL ENCODING

Type of delta encoding whereby common prefixes or suffixes and their lengths are recorded so that they need not be duplicated.
This works best with sorted data.

Original Data

rob
robbed
robbing
robot

Common Prefix

-
rob
robb
rob

INCREMENTAL ENCODING

Type of delta encoding whereby common prefixes or suffixes and their lengths are recorded so that they need not be duplicated.
This works best with sorted data.

Original Data

rob
robbed
robbing
robot

Common Prefix

-
rob
robb
rob

Compressed Data

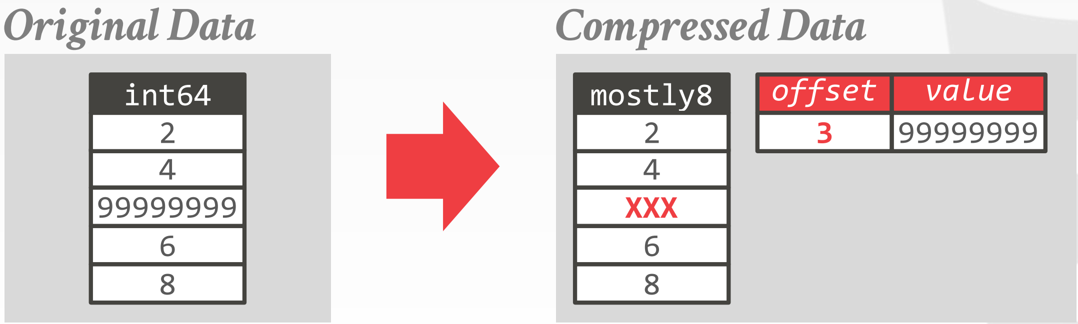
0	rob
3	bed
4	ing
3	ot

Prefix Length **Suffix**

MOSTLY ENCODING

When the values for an attribute are "mostly" less than the largest size, you can store them as a smaller data type.

→ The remaining values that cannot be compressed are stored in their raw form.



Source: Redshift Documentation

DICTIONARY COMPRESSION

Replace frequent patterns with smaller codes.

Most pervasive compression scheme in DBMSs.

Need to support fast encoding and decoding.

Need to also support range queries.

DICTIONARY-BASED ORDER-PRESERVING STRING COMPRESSION FOR MAIN MEMORY COLUMN STORES
SIGMOD 2009

DICTIONARY COMPRESSION

When to construct the dictionary?

What is the scope of the dictionary?

How do we allow for range queries?

How do we enable fast encoding/decoding?

DICTIONARY CONSTRUCTION

Choice #1: All At Once

→ Compute the dictionary for all the tuples at a given point of time.

→ New tuples must use a separate dictionary or the all tuples must be recomputed.

Choice #2: Incremental

→ Merge new tuples in with an existing dictionary.

→ Likely requires re-encoding to existing tuples.

DICTIONARY SCOPE

Choice #1: Block-level

→ Only include a subset of tuples within a single table.
→ Potentially lower compression ratio, but can add new tuples more easily.

Choice #2: Table-level

→ Construct a dictionary for the entire table.
→ Better compression ratio, but expensive to update.

Choice #3: Multi-Table

→ Can be either subset or entire tables.
→ Sometimes helps with joins and set operations.

MULTI-ATTRIBUTE ENCODING

Instead of storing a single value per dictionary entry, store entries that span attributes.
→ I'm not sure any DBMS actually implements this.

Original Data

val1	val2
A	202
B	101
A	202
C	101
B	101
A	202
C	101
B	101

MULTI-ATTRIBUTE ENCODING

Instead of storing a single value per dictionary entry, store entries that span attributes.

→ I'm not sure any DBMS actually implements this.

Original Data

val1	val2
A	202
B	101
A	202
C	101
B	101
A	202
C	101
B	101

Compressed Data

val1+val2
XX
YY
XX
ZZ
YY
XX
ZZ
YY

val1	val2	code
A	202	XX
B	101	YY
C	101	ZZ

ENCODING / DECODING

A dictionary needs to support two operations:

→ **Encode:** For a given uncompressed value, convert it into its compressed form.

→ **Decode:** For a given compressed value, convert it back into its original form.

No magic hash function will do this for us.

We need two data structures to support operations in both directions.

ORDER-PRESERVING COMPRESSION

The encoded values need to support sorting in the same order as original values.

Original Data

name
Andrea
Prashanth
Andy
Dana

ORDER-PRESERVING COMPRESSION

The encoded values need to support sorting in the same order as original values.

Original Data

name
Andrea
Prashanth
Andy
Dana

Compressed Data

name
10
40
20
30

value	code
Andrea	10
Andy	20
Dana	30
Prashanth	40

ORDER-PRESERVING COMPRESSION

The encoded values need to support sorting in the same order as original values.

```
SELECT * FROM users
 WHERE name LIKE 'And%'
```

```
SELECT * FROM users
 WHERE name BETWEEN 10 AND 20
```

Original Data

name
Andrea
Prashanth
Andy
Dana

Compressed Data

name
10
40
20
30

value	code
Andrea	10
Andy	20
Dana	30
Prashanth	40

ORDER-PRESERVING COMPRESSION

```
SELECT name FROM users
 WHERE name LIKE 'And%'
```

Still have to perform seq scan

Original Data

name
Andrea
Prashanth
Andy
Dana

Compressed Data

name
10
40
20
30

value	code
Andrea	10
Andy	20
Dana	30
Prashanth	40

ORDER-PRESERVING COMPRESSION

```
SELECT name FROM users
 WHERE name LIKE 'And%'
```

→ *Still have to perform seq scan*

```
SELECT DISTINCT name
   FROM users
 WHERE name LIKE 'And%'
```

→ *Only need to access dictionary*

Original Data

name
Andrea
Prashanth
Andy
Dana

→

Compressed Data

name
10
40
20
30

value	code
Andrea	10
Andy	20
Dana	30
Prashanth	40

DICTIONARY IMPLEMENTATIONS

Hash Table:

→ Fast and compact.

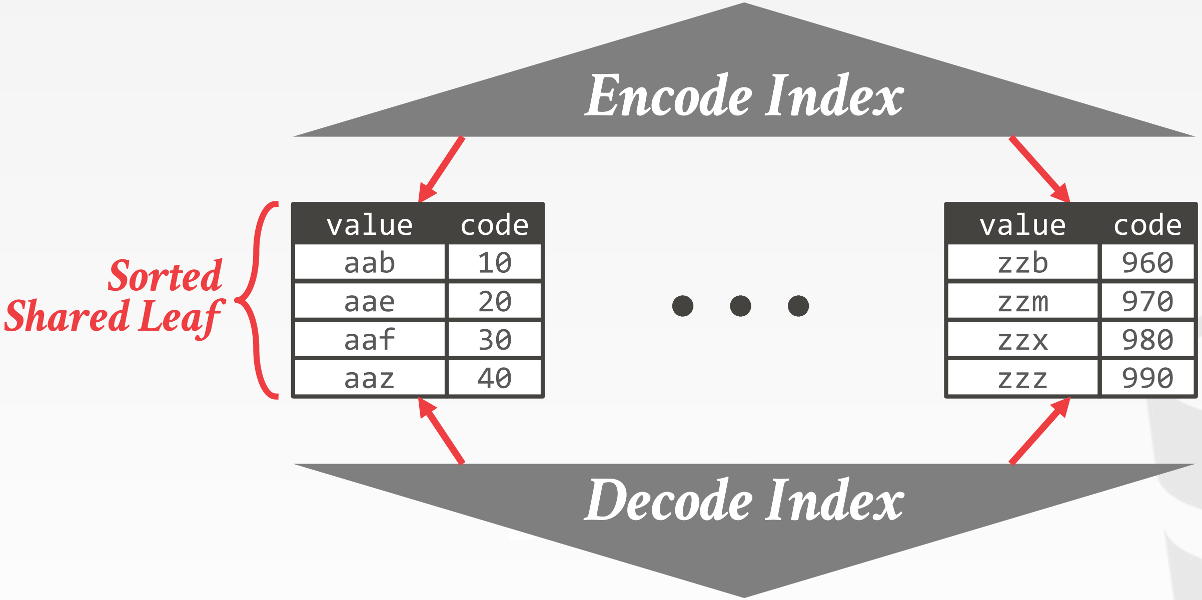
→ Unable to support range and prefix queries.

B+Tree:

→ Slower than a hash table and takes more memory.

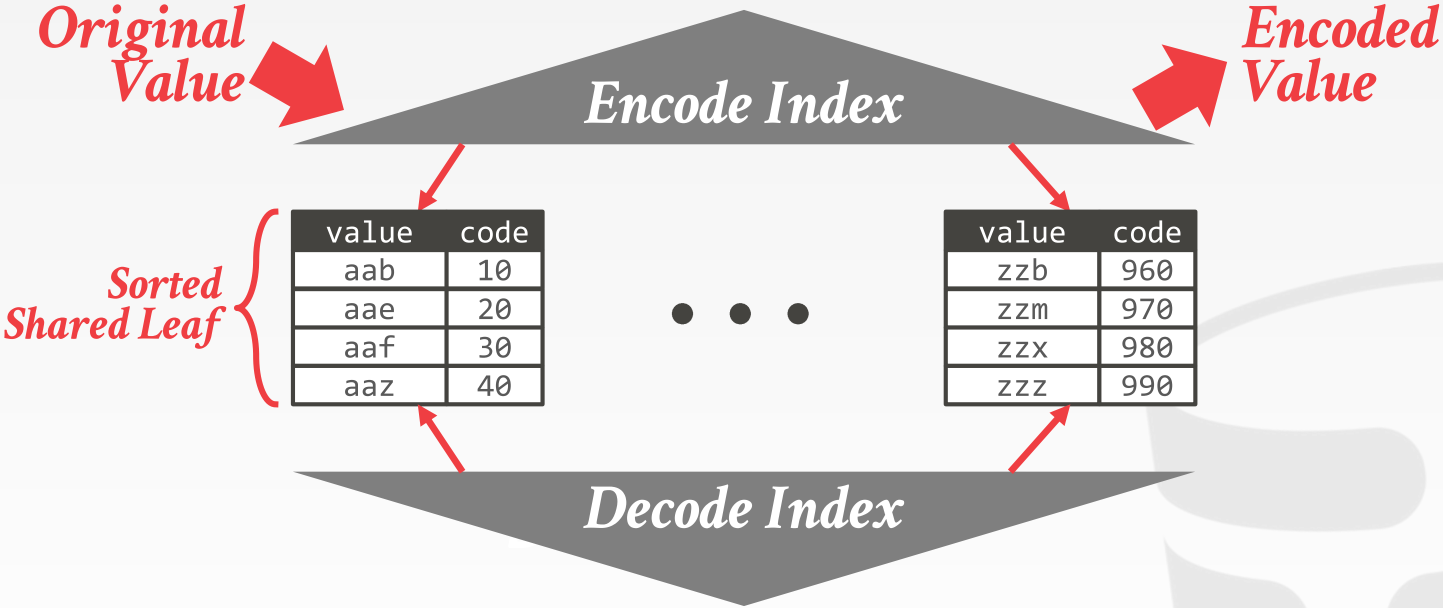
→ Can support range and prefix queries.

SHARED-LEAVES TREES



DICTIONARY-BASED ORDER-PRESERVING STRING COMPRESSION FOR MAIN MEMORY COLUMN STORES
SIGMOD 2009

SHARED-LEAVES TREES

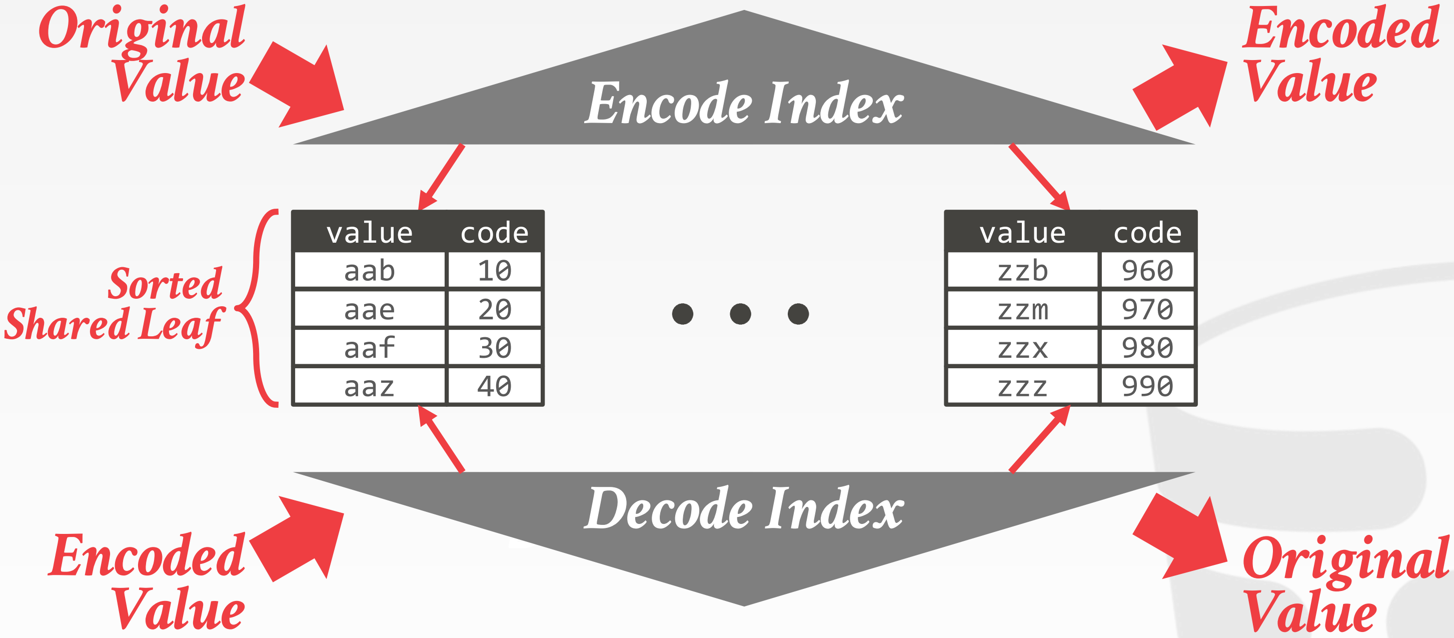


value	code
aab	10
aae	20
aaf	30
aaz	40

value	code
zzb	960
zzm	970
zzx	980
zzz	990

DICTIONARY-BASED ORDER-PRESERVING STRING COMPRESSION FOR MAIN MEMORY COLUMN STORES
SIGMOD 2009

SHARED-LEAVES TREES

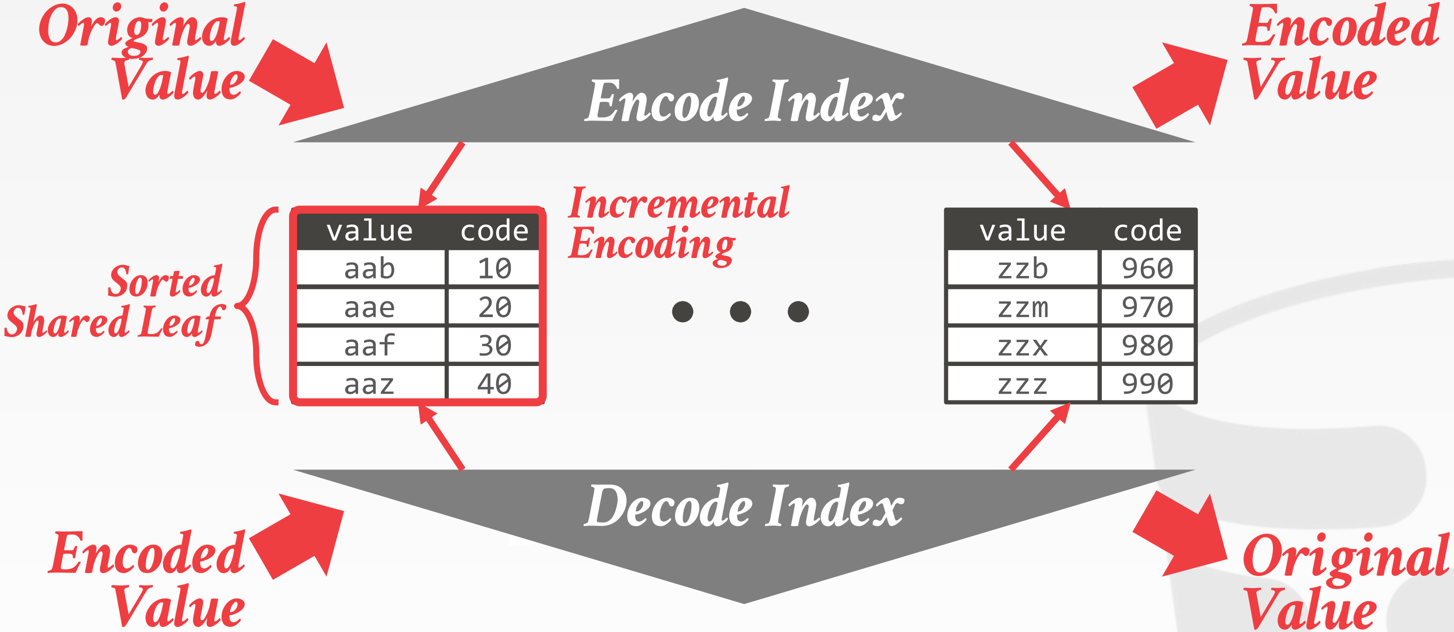


value	code
aab	10
aae	20
aaf	30
aaz	40

value	code
zzb	960
zzm	970
zzx	980
zzz	990

DICTIONARY-BASED ORDER-PRESERVING STRING COMPRESSION FOR MAIN MEMORY COLUMN STORES
SIGMOD 2009

SHARED-LEAVES TREES



value	code
aab	10
aae	20
aaf	30
aaz	40

• • •

value	code
zzb	960
zzm	970
zzx	980
zzz	990

DICTIONARY-BASED ORDER-PRESERVING STRING COMPRESSION FOR MAIN MEMORY COLUMN STORES
SIGMOD 2009

PARTING THOUGHTS

Dictionary encoding is probably the most useful compression scheme because it does not require pre-sorting.

The DBMS can combine different approaches for even better compression.

It is important to wait as long as possible during query execution to decompress data.

NEXT CLASS

Physical vs. Logical Logging