

Corra: Correlation-Aware Column Compression

Hanwen Liu
hanwen.liu@tum.de
TUM

Mihail Stoian
mihail.stoian@utn.de
UTN

Alexander van Renen
alexander.van.renen@utn.de
UTN

Andreas Kipf
andreas.kipf@utn.de
UTN

ABSTRACT

Column encoding schemes have witnessed a spark of interest with the rise of open storage formats (like Parquet) in data lakes in modern cloud deployments. This is not surprising – as data volume increases, it becomes more and more important to reduce storage cost on block storage (such as S3) as well as reduce memory pressure in multi-tenant in-memory buffers of cloud databases. However, single-column encoding schemes have reached a plateau in terms of the compression size they can achieve.

We argue that this is due to the neglect of cross-column correlations. For instance, consider the column pair (*city*, *zip_code*). Typically, cities have only a few dozen unique zip codes. If this information is properly exploited, it can significantly reduce the space consumption of the latter column.

In this work, we depart from the established path of compressing data using only single-column encoding schemes and introduce several what we call *horizontal*, correlation-aware encoding schemes. We demonstrate their advantages over single-column encoding schemes on the well-known TPC-H’s *lineitem*, LDBC’s *message*, *DMV*, and *Taxi* datasets. Our correlation-aware encoding schemes save up to 58.3% of the compressed size over single-column schemes for *lineitem*’s *receiptdate*, 53.7% for *DMV*’s *zip_code*, and 85.16% for *Taxi*’s *total_amount*.

CCS CONCEPTS

• **Information systems** → **Data compression**.

KEYWORDS

column correlation, column encoding schemes, data compression

VLDB Workshop Reference Format:

Hanwen Liu, Mihail Stoian, Alexander van Renen, and Andreas Kipf. Corra: Correlation-Aware Column Compression. VLDB 2024 Workshop: 2nd Workshop on Cloud Databases.

1 INTRODUCTION

Column encoding schemes lie at the heart of the storage layer of any database system. There are several single-column encoding schemes that are by now already ad-hoc: Frame-of-Reference (FOR), Frequency, Dictionary, Delta, Run-Length Encoding (RLE), FSST [5], among many others (see, e.g., Lemire and Boytsov [12]). However, data in the wild is highly correlated and the current encoding schemes do not exploit this fact. Only recently Lyu et al. [15] made

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.
Proceedings of the VLDB Endowment. ISSN 2150-8097.

shipdate	commitdate	shipdate	commitdate
1992-01-02	1992-03-10	1992-01-02	<--- +68
1998-12-01	+2369	1998-12-01	<--- -88
2024-06-08	+9417	2024-06-08	<--- +8

(a) Vertical encodings (prior) (b) Horizontal encodings (ours)

Figure 1: Vertical (prior work) vs. horizontal encodings (ours): Exploiting the correlation between date columns in TPC-H’s *lineitem* table for the column pair (*shipdate*, *commitdate*).¹ Instead of encoding *commitdate* w.r.t. to its own values (vertically), it is better to encode it w.r.t. *shipdate* (horizontally). The dashed arrows show the corresponding dependency.

the case for correlation-aware *bitmaps*, showing that one can indeed lower the space requirement while maintaining a decent runtime overhead during decompression.

In this work, we propose CORRA, a unified collection of novel *horizontal*, correlation-aware column encoding schemes that reduce the compressed size beyond what is possible with current single-column encoding schemes, while maintaining negligible query runtime overhead. This is in contrast to previous work that relies on *vertical* column encoding schemes, i.e., limited to the encoded column alone, as we argue next.

Related Work. There is an extensive line of research focusing on single-column encoding schemes, with a recent spark of attention with works such as BtrBlocks [9] and FastLanes [3]. With BtrBlocks, Kuschewski et al. [9] argue that open formats such as Apache’s Parquet tend to be rather inefficient in terms of decompression time, resulting in scans begin CPU-bound, thus increasing query time. To address this, they employ several encoding schemes for different data types, which can also be applied recursively. FastLanes [3] enables efficient vectorized execution by cleverly reordering tuples to maximize the effect of SIMD operations. BitWeaving [13] takes a different approach by interleaving the bits of the column values. Orthogonal to these, FSST [5] and Fast & Strong [11] specialize in improved string compression, while other works focus on floating-point compression [4, 14]; in particular, these are orthogonal to our work since we do not constrain the encoding scheme of the reference column (see the upcoming section for a definition).

Trummer [19] investigates whether LLMs can infer correlations from the schema itself. Correlations have been considered for indexes to improve query processing, namely Kimura et al. [8] introduce the Correlation Map as a secondary index, concept later generalized in Cortex [16]. Instead, we exploit correlations for data compression. We thus depart from the established, well-trodden path of compressing data using only single-column encoding schemes and introduce *correlation-aware* encoding schemes that push the

boundaries of what is possible with single-column coding schemes in terms of compression size.

2 HORIZONTAL ENCODING SCHEMES

The key idea behind horizontal column compression is to express the values of one column, which we call the *diff-encoded* column, in terms of the other, which we call the *reference* column. The intuition is that this reduces the range of values the diff-encoded values can take, so the bit-width required to compress the column decreases accordingly. In the following, we propose two correlation-aware encoding schemes that follow this paradigm, and present real-world datasets which these schemes are applicable to. In addition, we show that correlation-aware compression is not bound to using a single reference column and extend the diff-encoding paradigm to *multiple* reference columns.

2.1 Non-hierarchical Encoding

Let us take a closer look at the table `lineitem` of the well-known decision support benchmark TPC-H [18]: The table, which stores detailed information about each line item in a customer order, has three date-valued columns, corresponding to the three relevant dates in a supply chain: `shipdate`, `commitdate`, and `receiptdate`. In particular, it is a known fact that the differences between these dates are bounded, namely at most several months apart from each other. Such a correlation can be indeed exploited: Indeed, instead of storing the column `commitdate` as is, one can instead store the *difference* to `shipdate`, hence reducing the bit-width required to store it; similarly for `receiptdate`. To see why this is the case, consider Fig. 1, where we illustrate this case. We fix `shipdate` and `commitdate` as the reference and diff-encoded columns, respectively. Hence, we compute the difference of `commitdate` to `shipdate`. This results in smaller numbers in the `commitdate` column, thus reducing the necessary bit-width to store it.

Optimal Diff-Encoding. The careful reader may be interested in how we can actually configure which columns to diff-encode and which to compress using standard single-column compression. We use a strategy similar to the one used in CorBit [15]. To this end, consider Fig. 2, which visualizes our method.² The vertices represent the column, so for each directed edge $a \rightarrow b$, we measure what the compressed size would be if we diff-encoded a w.r.t. b . Using a cost-based greedy strategy, we can then decide which columns will be reference columns and which will not. In our case, `l_shipdate` is the reference column for both `l_commitdate` and `l_receiptdate`, saving 82.5 MB over just bit-packing the individual columns (TPC-H, SF 10); the exact numbers can be found in Tab. 2. Naturally, considering cases where a diff-encoded column becomes itself a reference column is an interesting future work (we did not focus on this type of diff-encoding configuration).

Outlier Detection. Note that our scheme is beneficial when there is a *bounded* difference in the diff-encoded column. In the case where the differences are not always bounded, we also design and implement an outlier storage architecture. This will be particularly useful when extending to the case of multiple reference columns

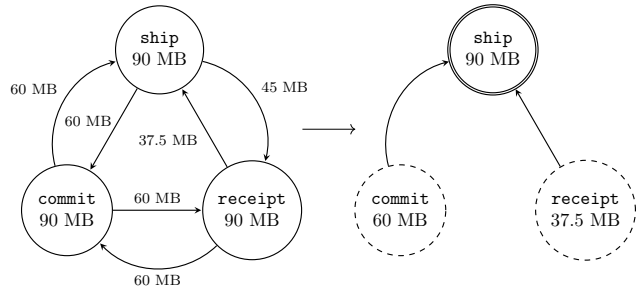


Figure 2: Detecting the optimal diff-encoding configuration in TPC-H (SF 10) for its three date-valued columns. The weight of an $a \rightarrow b$ edge is the size of column a when diff-encoded w.r.t. reference column b .

city	zip-code	city	zip-code
Cortland	13045	0	0
Naples	34102	1	0
Naples	34112	1	1
Naples	34102	1	0
NYC	10016	2	0
NYC	10001	2	1

(a) Without diff-encoding

(b) With diff-encoding

Metadata for hierarchical encoding:

```
city_dict: {0 : "Cortland", 1 : "Naples", 2 : "NYC"}
zip_codes: [13045, 34102, 34112, 10016, 10001]
offsets: [0, 1, 3]
```

Figure 3: Hierarchical encoding: Exploiting the correlation of the column-pair (city, zip-code) in the DMV dataset. The metadata contains an array of zip-codes along with an array of offsets for each individual city starting from. The city dictionary is used to reconstruct the city column.

(discussed in Sec. 2.3). Notably, in the datasets we evaluated on, the simple case of single reference columns did not require any special outlier handling.

2.2 Hierarchical Encoding

The next correlation-aware encoding scheme we introduce is *hierarchical* encoding, targeting column pairs that are endowed with a natural hierarchical-like structure. Consider the column pair (`city`, `zip-code`) of the DMV dataset: While the `zip-code` column has many distinct values, a specific `city` only has a few. Intuitively, this reduces the bit-width necessary to store these zip-codes.

Example. In Fig. 3, we illustrate our compression scheme for the column pair (`city`, `zip-code`) of the DMV dataset [1]. The key insight is that a specific pair can occur many times in the dataset – we exemplify this using two records (Naples, 34102). Our compression scheme works as follows: We collect the different zip-codes of each city in a metadata array “`zip_codes`”. This array is indexed by each distinct city with an auxiliary “`offsets`” array, i.e., the zip-codes of

¹We add a custom row to our TPC-H sample.

²To avoid clutter, we round up the exact numbers in this figure.

Cortland start at position 0, those of Naples at position 1, while those of NYC start at position 3 (highlighted in color).

Compression. To obtain the metadata for this encoding scheme, we maintain a hashtable of cities on the fly and their corresponding zip-codes. The “zip_code” array and the “offsets” array can then be computed once the compression has been finalized, in a single pass as well.

Decompression. Naturally, as for the other compression scheme we discussed, to access a certain value within the zip-code column, we need to also access the corresponding city, as shown in Alg. 1: To decompress the zip-code value of tuple “tid”, we fetch both the city (which has been dict-encoded in advance) and access the array zip-codes at the offset *shifted* by the diff-encoded value of the zip-code column.

Algorithm 1 HIERARCHICALACCESS(tid)

```

1: ref ← FETCH(city)[tid]
2: diff ← FETCH(zip-code)[tid]
3: return zip_codes[offset[ref] + diff]

```

2.3 Supporting Multiple Reference Columns

We have shown that correlation-aware encoding schemes are able to improve upon correlation-agnostic ones. In this section, we explore how multiple reference columns can be used to unlock further compression opportunities. Particularly, we show that the non-hierarchical encoding scheme (Sec. 2.1) is not limited to using a *single* reference column. We show evidence that we can have as well *multiple* reference columns.

Example. Typically, the total amount in a dataset should directly equal the sum of its parts. However, in real-world data, this relationship is often more fuzzy. Therefore, we need smarter encoding schemes to handle data points that do not follow simple logic. Fortunately, most data adhere to identifiable arithmetic logic. We can encode these logics rather than the data itself, enabling compression of the diff-encoded column. We observed this correlation among the monetary columns in the Taxi dataset [2], which contains taxi rides in NYC over the course of a year. We categorized these columns into three distinct groups:

- Group A: mta_tax, fare_amount, extra, tolls_amount, tip_amount, improvement_surcharge,
- Group B: congestion_surcharge,
- Group C: airport_fee.

These three groups serve as reference columns, along with one diff-encoded (target) column (total_amount). Unfortunately, the target column does not always follow directly from the sum of all three reference groups. Instead, it can be calculated using the following arithmetic logic in Tab. 1.

In this case, we can use 2-bit encoding (00, 01, 10, 11) to represent their correlation, indicating how to reconstruct from the reference columns. Additionally, we store a limited number of outliers (0.32%).

Compression. To efficiently store the encoding values and the outliers, we introduced an outlier storage design in CORRA, shown in Fig. 4. The values $\{V_1, V_2, V_3\}$ represent the original values, which

Group Representation	Probability	Binary Encoding
A	31.19%	00
A + B	62.44%	01
A + C	2.69%	10
A + B + C	3.33%	11
None	0.32%	outlier

Table 1: Diff-encoding column “total_amount” in the Taxi dataset [2] w.r.t. to multiple reference columns (see §2.3). In this case, the reference columns are partitioned in three groups, symbolically represented with A, B, and C.

can be calculated using arithmetic methods, $\{E_1, E_2, E_3\}$ represent the corresponding 2-bit codes (see Tab. 1), and $\{O_1, O_2\}$ represent the original values that cannot be calculated using the aforementioned methods, i.e., the outliers. In our example, their row-indices $\{1, 2\}$ and original values $\{O_1, O_2\}$ are stored in the outlier region after encoding. This allows us to encode each element of the target column with fewer bits, thereby saving space. Furthermore, we store the store the positions (indexes) and actual values of the outliers in an additional outlier storage area.

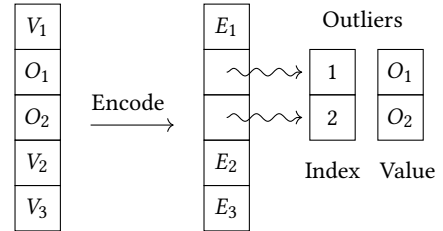


Figure 4: Non-hierarchical compression with multiple reference columns: Encoding the original target column with outliers (in this case, $\{O_1, O_2\}$). The regular values, $\{V_1, V_2, V_3\}$, are encoded as described in Tab. 1.

Decompression. During the decompression process, we first extract these two arrays from the outlier section to establish a mapping from outlier indexes to the outlier values. Subsequently, we reconstruct the encoded column element by element. We first use the established mapping for each position to check whether the element is an outlier. If it is an outlier, we directly assign the corresponding value. If it is not an outlier, we identify the corresponding encoding E to determine the computation method for reconstructing the target column from the reference columns. We then read the values from the reference columns to recover the original value.

We can also use a sentinel value to indicate the outlier values. However, this approach may introduce an additional value to our existing four encoding schemes, which would require using three bits to represent the encoding and cover all conditions. With our decompression design, the pre-stored outlier indexes determine if a specific decompression position is an outlier. There is no need to set a specific outlier encoding value at that position for distinction. The

Dataset	Column	Size w/o diff-enc	Encoding	Ref. column	Size w/ diff-enc	Saving rate
lineitem (SF 10)	l_receiptdate	89.99 MB	Non-hierarchical	l_shipdate	37.49 MB	58.3%
lineitem (SF 10)	l_commitdate	89.99 MB	Non-hierarchical	l_shipdate	59.99 MB	33.3%
Taxi	dropff	136.64 MB	Non-hierarchical	pickup	94.7 MB	30.6%
DMV	zip-code	25.88 MB	Hierarchical	city	11.96 MB	53.7%
DMV	city	21.45 MB	Hierarchical	state	21.05 MB	1.8%
message (SF 30)	ip	195.14 MB	Hierarchical	countryid	161.76 MB	17.1%
Taxi	total_amount	66.32 MB	Non-hierarchical	multiple (see §2.3)	9.84 MB	85.16%

Table 2: Space saving over single-column encoding schemes.

value can be any value from existing encoding values. Therefore, we can still use only two bits to indicate four types of arithmetic operations and outlier values.

3 EVALUATION

Datasets. In our evaluation, we utilize four datasets, two of which are synthetic:

- TPC-H’s `lineitem` [18]: We use scale factor 10, i.e., the table has 59,986,052 rows. We use its date-valued columns for the commit, ship, and receipt dates.
- LDBC’s `message` [6, 17]: We use SF 30, i.e., 76,388,857 rows.³ The table models a social network consisting of metadata related to messages that users post in discussion threads. We will use the highly correlated pair (`countryid`, `ip`) in our experiments.
- DMV [1]: The table of 12,176,621 rows consists of vehicle, snowmobile, and boat registrations in NYS.
- Taxi [2]: Taxi trips over one year (37,891,377 rows). We clean the dataset beforehand, i.e., remove rows where the drop-off happens *before* pickup, and remove the tuples where the money column is negative or out-of-range ($> 100\$$).

Experimental Setup. For ease of reproducibility, all experiments were performed on EC2. We used a `c5d.4xlarge` instance which has an Intel Xeon Platinum 8275CL processor with 16 vCPUs and 32 GB of memory.

We split all datasets into data blocks of 1M tuples. Each data block is completely self-contained: all information required to decompress it is contained within the block itself. When measuring query latency, we generate 10 uniform random selection vectors for each individual selectivity (as done, e.g., in Lang et al. [10]). In the experiment, we decompress and *materialize* the values at the specified positions, which we refer to as the query output.

Baseline. We compare CORRA to a baseline that employs the best single-column encoding scheme for each column. We use FOR- or Dict-encoding schemes, followed by a bit-packing. We chose these because they allow for fast random access into the compressed column; both RLE and Delta require checkpoints. To store column strings, we use Dict encoding and pack the distinct strings into a

flattened array. When it comes to query latency, we also consider the uncompressed case, i.e., no encoding schemes are used, so as to show the (minimal) overhead incurred by decoding.

Compression Size. We show that our correlation-aware encoding schemes are indeed superior to correlation-agnostic encoding schemes, which is the current status quo. In Tab. 2, we outline multiple columns within the four datasets for which our proposed correlation-aware column encoding schemes improve over the best configuration of single-column encoding schemes. In the case of hierarchical encoding, the compression size also includes the size of the metadata previously shown in Fig. 3.

Non-hierarchical Encoding. As already stated in Sec. 2.1, TPC-H’s `lineitem` table is a fertile ground for our non-hierarchical encoding scheme. Namely, we can diff-encode the date-valued columns `l_receiptdate` and `l_commitdate` w.r.t. `l_shipdate`, achieving saving rates of 58.3% and 33.3%, respectively. The difference between these rates arises because the time difference between the receiptdate and shipdate is smaller, allowing CORRA to apply shorter encodings, which results in a higher saving rate. Another example is the column-pair (`dropff`, `pickup`) in the Taxi dataset. This is a natural situation, since Taxi rides are short enough so that the difference between the two timestamps is not too large. For this, we obtain a saving rate of 30.6%.

Hierarchical Encoding. For our initial example in the DMV dataset, the column-pair (`city`, `zip-code`), we indeed observe a saving rate of 53.7%. A similar correlation can also be observed between `state` and `city`, yet with a lower saving rate, because even if we save a lot by reducing the bit-width in the element encoding, we still need to store all distinct strings. Such hierarchical correlations are also present in LDBC’s `message` table: The number of IPs can be restricted to a specific country, thus reducing the necessary bit-width for storing the unique IPs via a dict-encoding. Using hierarchical encoding, we obtain a saving rate of 17.1% for this case.

Multiple Reference Columns. Using multiple reference columns also seems to be worth the effort. The size of `total_amount` column in the Taxi dataset is greatly reduced by our encoding, as its saving rate of 85.16% also reflects.

³Download link: <https://db.in.tum.de/~birler/data/ldbcbi-sf30.zip>

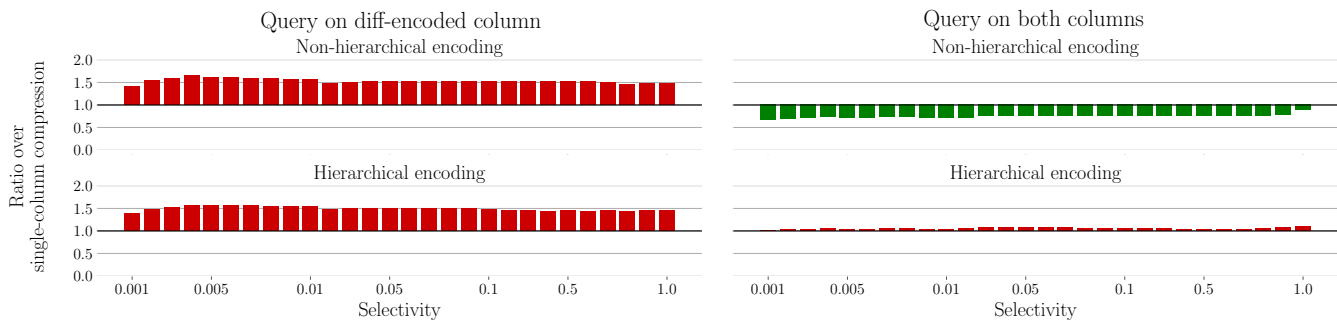


Figure 5: Query latency for selectivities in $\{0.001, 0.002, \dots, 0.9, 1.0\}$ with materialization of the query output. We run non-hierarchical encoding (§2.1) on TPC-H’s *lineitem* (SF 10) for *l_shipdate* (reference) and *l_receiptdate* (diff-encoded), and hierarchical encoding (§2.2) on LDBC’s *message* (SF 30) for *countryid* (reference) and *ip* (diff-encoded).

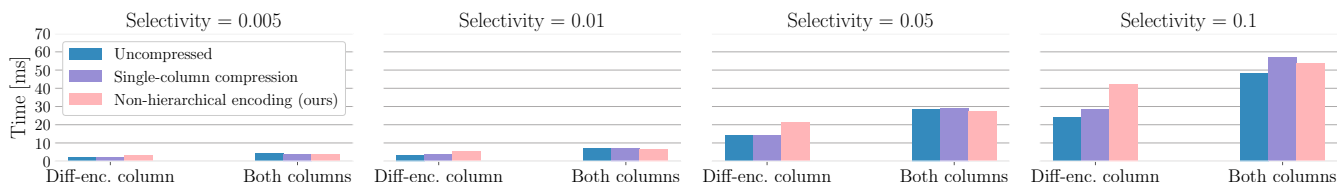


Figure 6: Non-hierarchical encoding (§2.1) Zooming in for different selectivities, including the “uncompressed” case.

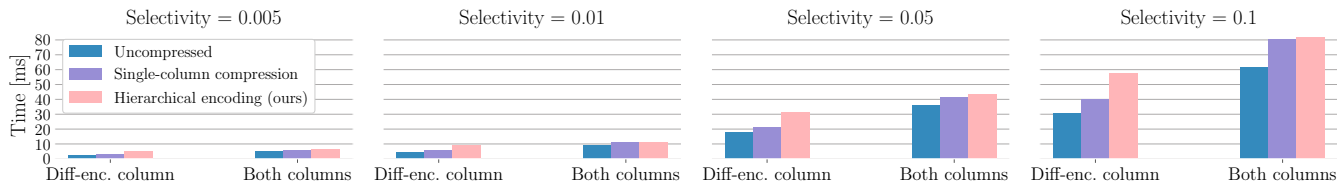


Figure 7: Hierarchical encoding (§2.2): Zooming in for different selectivities, including the “uncompressed” case.

Query Latency. When fetching the target column, CORRA must first fetch the reference column, which results in additional overhead. In Fig. 5, we show the query latency for different selectivities in the set $\{0.001, 0.002, \dots, 0.9, 1.0\}$ when querying (i) the diff-encoded column, and (ii) both columns.

Non-hierarchical Encoding. We show the query latency for the column pair (*l_shipdate*, *l_commi tdate*) – as before, *l_shipdate* is the reference column, while *l_commi tdate* is the diff-encoded column. We show the slow-down / speed-up compared to single-column encoding schemes. Querying the diff-encoded column results in a maximum slow-down of 1.66x, while querying *both* columns gives us an advantage since we have to access the reference column anyway.

Hierarchical Encoding. We evaluate the column pair (*countryid*, *ip*). The *ip* is subordinate to the *countryid*. This is a typical hierarchical encoding scenario. Considering the high repetition rate in the *ip* column, baseline compression uses dictionary encoding for the *ip* column. The query latency graph shows that when we only query the target, diff-encoded column, the slowdown trend is similar to that in non-hierarchical encoding. The slowdown ranges

from 1.39x to 1.56x. When querying both columns simultaneously, there is almost no slowdown over vertical encoding schemes.

Let us compare the two encoding schemes based on the query latency of fetching both columns. Hierarchical encoding slows down more than non-hierarchical encoding. This is because hierarchical encoding requires a complex reconstruction procedure, including more memory access. Non-hierarchical encoding reconstructs the second column by direct addition. This difference results in a different speedup/slowdown effect in these two encoding schemes.

Multiple Reference Columns. In Fig. 8, we show the query latency for non-hierarchical compression with multiple reference columns on the *Taxi* dataset. As discussed in Sec. 2.3, the target column, *total_amount*, is based on eight reference columns. Reconstructing the target column requires fetching and computing based on all reference columns. Fetched data is typically scattered at low selectivities, resulting in lower cache hit rates, and thus, the slowdown ratio remains high. As the volume of queried data increases, data locality improves, leading to a gradual decrease in the slowdown ratio, stabilizing around the 2x range. During full-range queries, the slowdown ratio exhibits a slight increase due to the handling of all

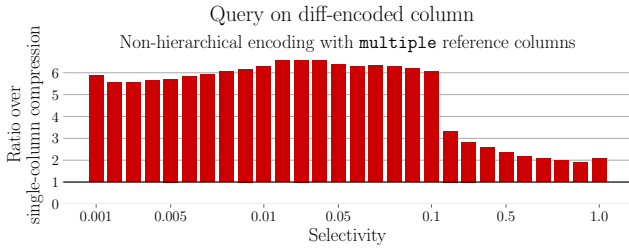


Figure 8: Query latency for non-hierarchical compression with multiple reference columns, specifically eight (cf. §2.3). The queries are run on Taxi dataset with eight reference columns and the diff-encoded “total_amount”.

outliers. The additional processing required for outliers contributes to the increase in average latency.

Zoom-In. In Fig. 6, we provide a zoom-in for the query latency in the case of non-hierarchical encoding, namely, we plot the query latency for four different selectivities {0.005, 0.01, 0.05, 0.1}. Additionally, we also consider the “uncompressed” case in which the query is directly executed over the uncompressed column(s). The overhead of our approach can be seen when considering the query latencies for the diff-encoded column. However, this overhead is mitigated when querying both columns, since in this case we need to read both columns anyway.

Similarly, for hierarchical encoding, in Fig. 7, we zoom in on the latency plot for the same four selectivities. The query suffers a small overhead in that we need the (un-prefetchable) lookup into the “locationips”-array (compared to the “zip_codes”-array in Fig. 3). In this case, this incurred overhead is not completely mitigated when querying both columns, as was the case for non-hierarchical encoding (in non-hierarchical encoding, there is no additional metadata).

Independent Work. Independently of our work, Glas et al. [7] also proposed correlation-aware column encoding schemes: They introduce C3, which further improves the compression sizes of BtrBlocks [9]. Their encoding schemes bear similarity to ours: They further generalize the implementation of the non-hierarchical encoding scheme as an affine function and explore more implementations of hierarchical encoding schemes, e.g., using FOR for the diff-encoded column. C3 also focuses on the optimal diff-encoding scheme, similar to our proposed selection scheme (see Fig. 2).

However, Glas et al. [7] do not consider diff-encoding with *multiple* reference columns, as we discuss in Sec. 2.3. In addition, they do not benchmark the query latency on their compressed columns, as we extensively did for our encoding schemes; see Fig. 5 and Fig. 8. In the following, we compare CORRA to C3, where we let C3 choose the (correlation-aware) encoding scheme for a given pair of columns. In Tab. 3, we summarize the achieved saving rates. The main observation is that CORRA and C3 are on par in terms of compression size. The only exception is the (pickup, dropff) pair in the Taxi dataset, where C3’s numerical encoding scheme manages to further exploit the (affine-like) correlation between the two timestamps.

Column-Pair	CORRA (ours)	C3 [7]
(shipdate, commitdate)	33.3% (§2.1)	31.5% (DFOR)
(shipdate, receiptdate)	58.3% (§2.1)	56.1% (DFOR)
(pickup, dropff)	30.6% (§2.1)	52.9% (Numerical)
(city, zip-code)	53.7% (§2.2)	59.1% (1-to-1)

Table 3: Saving rates on our datasets compared to the independent work by Glas [7], C3. Their DFOR is a hierarchical encoding where the diff-encoded column is compressed via FOR, the numerical encoding scheme generalizes the non-hierarchical encoding scheme as an affine function, and their 1-to-1 is specialized for the case where one could directly infer the diff-encoded column from the reference column. Also, C3 does not support multiple reference columns (§2.3).

4 CONCLUSION

Vertical, single-column encoding schemes have reached a plateau in terms of compression size. We deviate from this path and propose horizontal, correlation-aware encoding schemes. These achieve compression sizes beyond what is currently possible with standard, single-column schemes. Our *non-hierarchical* and *hierarchical* encoding schemes emulate natural column correlations that can be found in established datasets. The key idea behind them is to *diff-encode* a column with respect to one or more reference columns.

Consequently, we have shown that CORRA achieves significant reductions in the compressed size of several columns of TPC-H’s `lineitem`, LDBC’s `message`, `DMV`, and `Taxi`. Additionally, we provided a strategy to find an optimal diff-encoding configuration, i.e., which columns to diff-encode and which to compress via standard single-column encoding schemes. We envision CORRA to support more fine-grained column-aware correlation types and, particularly, automatic correlation detection, especially for our non-hierarchical encoding scheme with multiple reference columns.

REFERENCES

- [1] 2020. *Vehicle, snowmobile, and boat registrations*. <https://catalog.data.gov/dataset/vehicle-snowmobile-and-boat-registrations>
- [2] 2024. *Yellow Taxi trip records*. <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
- [3] Azim Afrozeh and Peter A. Boncz. 2023. The FastLanes Compression Layout: Decoding >100 Billion Integers per Second with Scalar Code. *Proc. VLDB Endow.* 16, 9 (2023), 2132–2144. <https://doi.org/10.14778/3598581.3598587>
- [4] Azim Afrozeh, Leonardo X. Kuffo, and Peter A. Boncz. 2023. ALP: Adaptive Lossless floating-Point Compression. *Proc. ACM Manag. Data* 1, 4 (2023), 230:1–230:26. <https://doi.org/10.1145/3626717>
- [5] Peter A. Boncz, Thomas Neumann, and Viktor Leis. 2020. FSST: Fast Random Access String Compression. *Proc. VLDB Endow.* 13, 11 (2020), 2649–2661. <http://www.vldb.org/pvldb/vol13/p2649-boncz.pdf>
- [6] Orri Erling, Alex Averbuch, Josep Lluís Larriba-Pey, Hassan Chafi, Andrey Gubichev, Arnau Prat-Pérez, Minh-Duc Pham, and Peter A. Boncz. 2015. The LDBC Social Network Benchmark: Interactive Workload. In *SIGMOD Conference*. ACM, 619–630.
- [7] Thomas Glass. 2023. Exploiting Column Correlations for Compression. <https://homepages.cwi.nl/~boncz/>.
- [8] Hideaki Kimura, George Huo, Alexander Rasin, Samuel Madden, and Stanley B. Zdonik. 2009. Correlation Maps: A Compressed Access Method for Exploiting Soft Functional Dependencies. *Proc. VLDB Endow.* 2, 1 (2009), 1222–1233. <https://doi.org/10.14778/1687627.1687765>

- [9] Maximilian Kuschewski, David Sauerwein, Adnan Alhomssi, and Viktor Leis. 2023. BtrBlocks: Efficient Columnar Compression for Data Lakes. *Proc. ACM Manag. Data* 1, 2 (2023), 118:1–118:26. <https://doi.org/10.1145/3589263>
- [10] Harald Lang, Alexander Beischl, Viktor Leis, Peter A. Boncz, Thomas Neumann, and Alfons Kemper. 2020. Tree-Encoded Bitmaps. In *SIGMOD Conference*. ACM, 937–967.
- [11] Robert Lasch, Ismail Oukid, Roman Dementiev, Norman May, Süleyman Sirri Demirsoy, and Kai-Uwe Sattler. 2019. Fast & Strong: The Case of Compressed String Dictionaries on Modern CPUs. In *DaMoN*. ACM, 4:1–4:10.
- [12] Daniel Lemire and Leonid Boytsov. 2015. Decoding billions of integers per second through vectorization. *Softw. Pract. Exp.* 45, 1 (2015), 1–29. <https://doi.org/10.1002/SPE.2203>
- [13] Yinan Li and Jignesh M. Patel. 2013. BitWeaving: fast scans for main memory data processing. In *SIGMOD Conference*. ACM, 289–300.
- [14] Panagiotis Liakos, Katia Papakonstantinou, and Yannis Kotidis. 2022. Chimp: Efficient Lossless Floating Point Compression for Time Series Databases. *Proc. VLDB Endow.* 15, 11 (2022), 3058–3070.
- [15] Xi Lyu, Andreas Kipf, Pascal Pfeil, Dominik Horn, Jana Giceva, and Tim Kraska. 2023. CorBit: Leveraging Correlations for Compressing Bitmap Indexes. In *VLDB Workshops (CEUR Workshop Proceedings)*, Vol. 3462. CEUR-WS.org.
- [16] Vikram Nathan, Jialin Ding, Tim Kraska, and Mohammad Alizadeh. 2020. Cortex: Harnessing Correlations to Boost Query Performance. *CoRR* abs/2012.06683 (2020). arXiv:2012.06683 <https://arxiv.org/abs/2012.06683>
- [17] Gábor Szárnyas, Jack Waudby, Benjamin A. Steer, Dávid Szakállas, Altan Birlir, Mingxi Wu, Yuchen Zhang, and Peter A. Boncz. 2022. The LDBC Social Network Benchmark: Business Intelligence Workload. *Proc. VLDB Endow.* 16, 4 (2022), 877–890. <https://doi.org/10.14778/3574245.3574270>
- [18] Transaction Processing Performance Council (TPC). 2022. TPC BENCHMARK™ H Standard Specification Revision 3.0.1. https://www.tpc.org/TPC_Documents_Current_Versions/pdf/TPC-H_v3.0.1.pdf. [Accessed 28-11-2023].
- [19] Immanuel Trummer. 2023. Can Large Language Models Predict Data Correlations from Column Names? *Proc. VLDB Endow.* 16, 13 (2023), 4310–4323. <https://www.vldb.org/pvldb/vol16/p4310-trummer.pdf>