# LDI: Learned Distribution Index for Column Stores

Dai-Hai Ton That, Mohammadsaleh Gharehdaghi, Alexander Rasin, Tanu Malik
School of Computing
DePaul University
Chicago, IL, USA
{dtonthat,mgharehd,arasin,tanu.malik}@depaul.edu

## ABSTRACT

In column stores, which ingest large amounts of data into multiple column groups, query performance deteriorates. Commercial column stores use log-structured merge (LSM) tree on projections to ingest data rapidly. LSM improves ingestion performance, but in column stores the sort-merge phase is I/O-intensive, which slows concurrent queries and reduces overall throughput. In this paper, we aim to reduce the sorting and merging cost that arise when data is ingested in column stores. We present LDI, a learned distribution index for column stores. LDI learns a frequency-based data distribution and constructs a *bucket* worth of data based on the learned distribution. Filled buckets that conform to the distribution are written out to disk; unfilled buckets are retained to achieve the desired level of sortedness, thus avoiding the expensive sort-merge phase. We present an algorithm to learn and adapt to distributions, and a robust implementation that takes advantage of disk parallelism. We compare LDI with LSM and production columnar stores using real and synthetic datasets.

## KEYWORDS

Learned Distribution Index, column-oriented, write-optimized, approximate clustering

## 1 INTRODUCTION

Column-oriented databases are a dominant backend DBMSes for supporting business decision-making processes [10, 36]. Column-stores, unlike their row-store counterparts, store entire columns contiguously, often in compressed form. Applications using column-oriented databases, typically, coalesce columns into groups. Using column groups significantly reduces the amount of data to be read, achieving high read performance for analytic (range-query) workloads in which most queries reference a column group.

In the era of big data, applications also ingest high volume data, often, arriving at high velocity. Log-structured merge tree (LSM-tree) logs incoming data in a buffer and periodically sort-merges the data [27, 32] into larger sorted runs. Typically used in wide-column NoSQL databases [2, 8, 18, 22], LSM-trees are increasingly available in column-store databases for fast writes and high throughput. Each group of columns in a column-store requires storage maintenance, thus column-stores have a greater need for a write-optimized index than row-stores. However, using an LSM index structure, which itself has a significant write amplification in a column-store database can also reduce query performance.

In a commercial system, such as Vertica [23, 33], all column groups must be indexed with an LSM-tree to preserve the row order in accordance with the primary key of the groups. Thus a write to a column group must coordinate with writes in other associated column groups for consistency. Queries, however, are often not uniform across column groups. Some column groups are more popular and queried more often than other less popular groups. Query performance of popular groups rapidly deteriorates due to concurrent writes across all groups, which defeats the purpose of dividing them into groups in the first place, and is one of the most important reasons why column-oriented databases are used.

One naive strategy is to split an insert into multiple individual inserts on column groups; this, however, will forsake the consistency and row order between column groups. An alternate strategy is to optimize the expensive sort-merge phase of LSM-trees. The key idea is that a fewer I/Os during sort-merge will lead to improved query performance and maintain consistency.

Current methods optimize this phase by adding summary structures within the buffer [7], improving when to merge [13, 14], and by measuring overlaps between buffer and on-disk data [4]. In every proposed approach, however, the sort-merge phase sorts *all* key values at periodic intervals of time. We show, analytically and experimentally, that this complete sorting of keys causes a large fraction of the I/O in an LSM-tree. In column stores this increase in I/O during inserts, reduces concurrent query performance, but, more importantly, this I/O due to sorting is redundant for answering analytical workloads. Our strategy, thus, is to eliminate the sort-merge phase of an LSM-tree and sort the incoming data *approximately*. The advantage of approximate sorting is that unlike LSM we do not need to wait for periodic intervals to merge and can write out incoming data as fast as it arrives; the disadvantage is that we must know in advance if the data being written out is *sufficiently* sorted.

In this paper, we present LDI, a low cost index for column stores, which logs data based on a learned distribution of data. If the incoming data conforms to the learned distribution, and the distribution remains stable, no further sorting will be needed for logged disk blocks. If the incoming data does not conform to the distribution, disk blocks will be approximately sorted and will need to be reorganized. We show that for real datasets, such a strategy localizes the reorganization instead of performing an entire sort-merge as in an LSM-tree.

LDI constructs a distribution similar to a dynamic histogram. We present an learning algorithm that decides when to adjust the intervals based on incoming data, and show how interval counts can


*SSDBM 2021, July 6-7, 2021, Tampa, Florida, USA*

ACM ISBN XXX-X-XXXXX-XXX-X.

| Order | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Key** | 9 | 10 | 3 | 18 | 1 | 13 | 20 | 24 | 12 | 4 | 2 | 7 | 19 | 30 | 26 | 14 | 1 | 29 | 8 | 4 | 15 | 25 | 18 | 9 | 2 | 8 | 11 | 6 | 16 | 5 | 12 | 28 |

**Figure 1: Sample data $D$ to insert.**

be maintained incrementally. LDI has the advantage of writing disk blocks as soon as data arrives. We present an I/O handler that takes advantage of in-built parallelism in HDD and SSD storage devices to support such a strategy. Finally we present an extensive set of experimental results comparing LDI with LSM and commercial column databases, using both real and synthetic datasets.

The rest of this paper is organized as follows. We present an example in Section 2. Section 2.1 analyses LSM-tree performance on column stores experimentally and analytically. We introduce learned distribution index structure (LDI) in Section 3, and its implementation details in Section 4. We show the experiments in Section 5, and discuss about the related work in Section 6. Finally, we conclude the paper in Section 7.

## 2 AN EXAMPLE: LSM VS LDI

In this section, we present an example to contrast LSM-Tree and LDI approach. There are two types of LSM trees, leveled (proactive merging) and tiered (delayed merging) [13, 21]. We illustrate leveled LSM tree in this example; tiered LSM tree example is presented in Appendix 8.1. We consider the costs of both tiered and leveled LSM trees throughout this paper.

Figure 1 lists a sequence $D$ of 32 incoming tuples (a tuple consists of a key and data, although only the key is shown) and the order in which they arrive. LSM tree periodically merges sorted runs of tuples into larger sorted runs either in the same level or in the higher level of the structure. Figure 2a shows the state of leveled LSM tree just before and just after a merge. The data arrives at $L0$, or the in-memory buffer with a size $B = 4$ (4 tuples). LSM is configured with a level ratio of $T = 3$ (representing the maximum size ratio between levels in LSM). If a merged level is full, LSM merges data runs at a higher level.

Figure 2a illustrates five merging steps performed by the LSM approach. The first four tuples (9, 10, 3, 18) are already sorted in memory and written to disk at $L1$ before step 1. In step 1, the sorted buffer (1, 13, 20, 24) is merged with $L1$ sorted run (3, 9, 10, 18) to write out the sorted sequence of eight tuples (1, 3, 9, 10, 13, 18, 20, 24) at $L1$. A step can trigger multiple consecutive merges: for example, in step 4, a merge between levels $L0$ and $L1$ leads to a full $L1$ level and is thus followed by a merge between levels $L1$ and $L2$. The ratio of data between levels is always maintained to be 3.

In cases where data is mostly sorted, LSM tree can choose to append sorted runs instead of merging them. For example, in step 5, a merge between levels $L0$ and $L1$ can be replaced by appending levels $L0$ and $L1$ instead. In this case, skipping the merge causes a small reduction in data sortedness (only tuple 5 is out of order). However, the likelihood of such near-sorted alignment depends on the distribution of the data. Although LSM can benefit from such distribution, it is not distribution-aware.

Intuitively, LDI designs the buffering process based on data distribution in order to ensure that data runs are mostly sorted and can be appended without incurring a merging operation. Figure 2b

| Step | Before Merge | After Merge |
|---|---|---|
| 1 | **L0** 1 13 20 24<br>L1 3 9 10 18 | 1 3 9 10 13 18 20 24 |
| 2 | **L0** 2 4 7 12<br>L1 1 3 9 10 13 18 20 24<br>L2 | 1 2 3 4 7 9 10 12 13 18 20 24 |
| 3 | **L0** 1 4 8 29<br>L1 14 19 26 30<br>L2 1 2 3 4 7 9 10 12 13 18 20 24 | 1 4 8 14 19 26 29 30<br>1 2 3 4 7 9 10 12 13 18 20 24 |
| 4 | **L0** 9 15 18 25<br>**L1** 1 4 8 14 19 26 29 30<br>L2 1 2 3 4 7 9 10 12 13 18 20 24 | 1 1 2 3 4 4 7 8 9 9 10 12<br>13 14 15 18 18 19 20 24 25 26 29 30 |
| 5 | **L0** 5 12 16 28<br>L1 2 6 8 11<br>L2 1 1 2 3 4 4 7 8 9 9 10 12<br>13 14 15 18 18 19 20 24 25 26 29 30 | 2 5 6 8 11 12 16 28<br>1 1 2 3 4 4 7 8 9 9 10 12<br>13 14 15 18 18 19 20 24 25 26 29 30 |

**(a) Inserting the sample data $D$ using leveled LSM trees.**

| Step | Buffer | Leaf intervals [1÷2] | [3÷6] | [7÷12] | [13÷19] | [20÷30] | Buckets written to disk |
|---|---|---|---|---|---|---|---|
| 1 | **9 10** 3 **18** | | (3) | (9,10) | (18) | | [9,10] |
| 2 | **1 13** 3 18 | (1) | (3) | | (18,13) | | [18,13] |
| 3 | 1 **20** 3 **24** | (1) | (3) | | | (20,24) | [20,24] |
| 4 | 1 **12** 3 **4** | (1) | (3,4) | (12) | | | [3,4] |
| 5 | 1 12 **2** **7** | (1,2) | | (7,12) | | | [1,2];[7,12] |
| 6 | **19 30 26 14** | | | | (14,19) | (26,30) | [14,19];[26,30] |
| 7 | **1 29 8 4** | (1) | (4) | (8) | | (29) | **[1,4]** |
| 8 | **15** 29 8 **25** | | | (8) | (15) | (29,25) | [29,25] |
| 9 | 15 **18** 8 **9** | | | (8,9) | (15,18) | | [8,9];[15,18] |
| 10 | **2 8 11 6** | (2) | (6) | (8,11) | | | [8,11] |
| 11 | 2 **16 5** 6 | (2) | (6,5) | | (16) | | [6,5] |
| 12 | 2 16 **12 28** | (2) | | (12) | (16) | (28) | **[12,16]** |
| 13 | 2 | (2) | | | | (28) | |

**(b) Inserting the sample data $D$ using LDI.**

**Figure 2: The behavior of LSM and LDI.**

illustrates LDI behavior for the data in Figure 1 with the same main buffer capacity $B = 4$. Leaf ranges are skewed based on the input data distribution. Each leaf bucket contains pointers to tuples stored in the buffer. For example in the first row buffer contains tuples (9, 10, 3, 8) and the interval $[3 \div 6]$ points to tuple 3, while $[13 \div 18]$ interval contains pointers for tuples 9 and 10.

The maximum number of pointers in a leaf interval is $n$ (in this example $n = 2$, and a bucket can hold 2 tuples). Every time an interval fills up, a new bucket is written to disk and cleared from the buffer. For instance, at step 1, the leaf interval $[7 \div 12]$ has two pointers (9 and 10). As a result, a new bucket $[9, 10]$ is created and written to disk. Tuples 9 and 10 are removed from the buffer and from the leaf storage. The last row in Figure 2b summarizes the buckets written during the ingestion of input data.

Unlike LSM-Tree, LDI runs do not require a merge as the distribution already writes nearly-sorted buckets. Table 1 shows the insertion cost of LSM-Tree and LDI in number of merges and the number of I/O operations. Leveled-LSM requires 6 merges with 36 writes and 20 reads; Tiered-LSM requires 2 merges with 28 writes and 8 reads. Meanwhile, LDI does not require any merges. The number of writes is equal to the total number of written buckets ($32/2 = 16$ buckets)

**Table 1: Merge cost with different methods**

| | Leveled-LSM | Tiered-LSM | LDI |
|---|---|---|---|
| #of merges | 6 | 2 | 0 |
| #of I/Os (Writes) | 36 I/O | 28 I/O | 16 I/O |
| #of I/Os (Reads) | 20 I/O | 16 I/O | 0 I/O |

Next, we examine read query performance using key-range queries. Without loss of generality $Q1$, $Q2$ and $Q3$ have the key-ranges [10, 12], [19, 20] and [14, 20], respectively. Table 2 presents the number of I/Os needed for each query. Tiered-LSM requires more I/O than both Leveled-LSM and LDI in all three queries. Leveled-LSM exhibits the best performance, but LDI query performance is equivalent for $Q1$ and $Q3$ and is only one I/O higher than Leveled-LSM for $Q2$.

**Table 2: Query cost with different methods**

| Range Query | Leveled-LSM | Tiered-LSM | LDI |
|---|---|---|---|
| $Q1 : [19, 20]$ | 2 I/O | 2 I/O | 2 I/O |
| $Q2 : [11, 12]$ | 2 I/O | 3 I/O | 3 I/O |
| $Q3 : [14, 20]$ | 4 I/O | 5 I/O | 4 I/O |

LDI data buckets can overlap because data is not strictly sorted and does not use merges. Therefore, some queries will read extraneous data at a higher cost. Ultimately, the goal of LDI design is to minimize these extra penalties by achieving good bucket compactness.

## 2.1 An Experiment: LSM Vs LDI in Column Databases

In the previous example we compared the performance of LSM trees with LDI on a small example data. In this section, we analyze the merge cost of LSM trees and measure it experimentally on columnar databases.

**Columnar databases.** Columnar databases store data tables by column where each column is stored separately. This allows a query to access that precise data that it needs. In general, each column can be stored separately, but this leads to high tuple reconstruction cost. Column grouping (or projection in C-Store [33] or Vertica [23]) is one way to reduce the tuple reconstruction cost. The idea is to group a subset of columns together, to benefit query operations that accesses all these columns. This group of columns is called *projection* in C-Store [33] or Vertica [23]. Column stores trade storage for improved tuple reconstruction cost and query access. For instance, it is possible to replicate columns across projections as well as support a *superprojection* with all columns. We assume a simplified columnar store model in which there are partitioned projections with no replication of columns across projections, and no superprojections.

**LSM tree merge cost.** We analyze the cost of levelling and tiering merge policies in LSM trees, which have not been mathematically formulated before. Merge policies are recursively defined and are initiated based on a pre-defined ratio between two consecutive levels.

- The levelling merge policy sort-merges whenever a level is full with previous runs in the next higher level.
- The tiering policy, on the other hand, is delayed merging of runs; it appends whenever a level is full with previous runs in the next higher level but sort-merges the runs only if the next higher level is determined to be full.

**Table 3: Notations used in this paper**

| Parameter | Description |
|---|---|
| $N$ | The total number of data entries |
| $B$ | The buffer size (Level $L_0$) |
| $M = \lfloor N/B \rfloor$ | The total number of buffers |
| $T$ | The ratio between two consecutive levels |
| $L = \lfloor log_T M \rfloor + 1$ | The total number of levels |
| $m^V, m^R$ | The total number of merges in Leveled/Tiered-LSM |

$$m^V = \sum_{i=1}^{L-1} \left( \left\lfloor \frac{M}{T^{i-1}} \right\rfloor - \left\lceil \frac{M}{T^i} \right\rceil \right) \quad (1)$$

$$m^R = \sum_{i=1}^{L-1} \left\lfloor \frac{M}{T^i} \right\rfloor \quad (2)$$

Equation 1 and 2 present the number of merges of leveling and tiering merge policy as a truncated geometric series defined based on $T$, the ratio between two consecutive levels and $M$, which is the total number of buffers. Tiered-LSM (Equation 2) only merges the data whenever a level on disk is full, $\left\lfloor \frac{M}{T^i} \right\rfloor$ represents the total number of times level $i$ gets full for a given $N$ entries Meanwhile, Leveled-LSM merges at level $i$ whenever level $i-1$ (including $L0$) gets full and data in level $i$ is not empty. In Equation 1, $\left\lfloor \frac{M}{T^{i-1}} \right\rfloor$ shows the number of times data is flushed from level $i-1$ to level $i$ and $\left\lceil \frac{M}{T^i} \right\rceil$ shows the number of times the data at level $i$ is empty (no merge is needed in case of level $i$ is empty.

Clearly tiered LSM causes fewer merges than levelled LSM but in both the cases the number is dominated by the constant factors that are multiplied on a per level basis. These constants play a significant role in a column database as shown in Figure 3. In this experiment we measure the total number of merges with different size of data in a columnar store. The column store has two configurations: 5 projections with each projection has 3 columns, and 1 projection with all columns. As shown in this Figure, while the number of merges of single LSM-Tree is quite reasonable, those of multiple

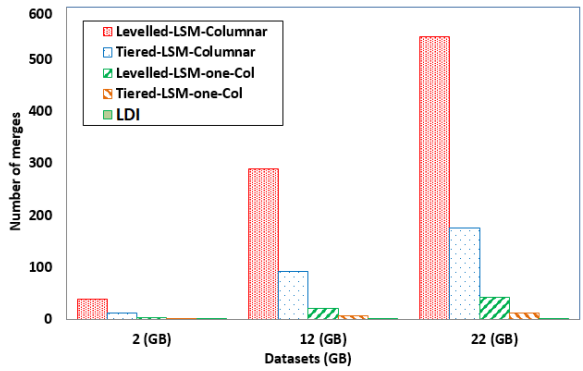


**Figure 3: The total number of merges during data-loading with tiered and levelled LSM trees. Multiple LSM trees on columnar stores, single LSM tree on one column and LDI. The columnar has 5 projections each of projection has 3 columns; whereas on-column keep the data in only one projection (all columns)**

LSM-Trees are multiplied by the factor of the number of projections and the $T$ factor. In contrast, a LDI has close to zero merges. We relate the number of merges to read/write costs in Appendix 8.1.

## 3 LEARNED DISTRIBUTION INDEX

We present LDI, our distribution-aware index, in this section. The basic idea in LDI is to use some fixed amount of training data to learn a given distribution, and then continuously update the learned distribution as new data comes in.

LDI is conservative in updating its distribution, since a change in the distribution will affect query costs. In order to learn, it maintains two distributions: a *global* distribution based on which buckets are created, and a *local* distribution, which reflects the state of incoming data. Local distribution may change more rapidly but bucket creation is determined by the global distribution. Only when sufficient evidence about the local distribution is collected, and a *drift* is detected, LDI updates the global distribution. In LDI a normalized frequency is used to create a distribution and is thus similar to V-optimal histograms but unlike dynamic V-optimal histograms [6], which keep tuning configurable, LDI decides when to tune.

We now describe in detail how the distribution is initialized and updated based on drifts. Because of differences between local and global distributions, LDI buckets are only approximately sorted. Since buckets are approximately sorted queries typically read more data than required. We present an I/O handler to improve query performance and a measure of sortedness which LDI provides to the user for undertaking, if necessary the clustering process (complete sorting of data).

### 3.1 Initializing a distribution

In LDI, a distribution is represented as an array of $n$ contiguous intervals $\{[b_i, b_{i+1})\}$, where $b_i$ and $b_{i+1}$ are interval boundaries. For an interval, we maintain two quantities: a normalized frequency, denoted $dis$ and a count of values denoted $load$.

These interval boundary values and frequency counts are initialized by inserting a fixed amount of data into a $k$-ary B$^+$-tree, and using the min/max key values in the leaf nodes as interval boundaries, and determining how many values fall in these leaf nodes. The value of $k$ of a B$^+$-tree determines the number of intervals we create. In the beginning, we decided to create a two level tree, and a high $k$ to create wider intervals.

### 3.2 Updating a distribution

Table 4 summarizes the variables maintained by LDI. Interval boundaries and frequency counts are updated as new data arrives. Local (denoted by $L$) and Current (denoted by $C$) have no recorded quantities (distribution or load) to begin with but its determines the frequency counts from a time window $t$ with $size$ data entries. For each time window $t$, we record the normalized frequency and the count of keys falling into each interval in $C$. Once time window $t$ is ended, the recorded normalized frequency $C.dis_i$, and the total count $C.load_i$ are accumulated. These local quantities are accumulated into local normalized frequency, denoted $L.dis_i$ and a local count of values denoted $L.load_i$.

We define the distribution drift$(L.dis_i, G.dis_i)$ as the ratio between the weighted estimate of the global normalized frequency and the local normalized frequency for the i$^{th}$ interval. Intuitively, a ratio of 1 represents that the data distribution has not changed and deviation from 1 represents an increase or decrease in the amount of data seen for this interval. If the drift is positive, more data is arriving, and we might be creating buckets that are too narrow; while if the drift is negative, less data is arriving, and we might be creating buckets that are too wide.

$$\text{drift}(i) = \frac{\text{w}_{L_i} * L.dis_i + \text{w}_{G_i} * G.dis_i}{G.dis_i} \tag{3}$$

**Table 4: Notation used in this paper**

| Parameter | Description |
|---|---|
| $G.b_i$ | Global boundary at interval $i^{th}$ |
| $G.dis_i$ | Global normalized frequency at interval $i^{th}$ |
| $G.load_i$ | Global count at interval $i^{th}$ |
| $L.b_i$ | Local boundary at interval $i^{th}$ |
| $L.dis_i$ | Local normalized frequency at interval $i^{th}$ |
| $L.load_i$ | Local count at interval $i^{th}$ |
| $C.b_i$ | Current boundary at interval $i^{th}$ |
| $C.dis_i$ | Current normalized frequency at interval $i^{th}$ |
| $C.load_i$ | Current count at interval $i^{th}$ |
| drift$(L.dis_i, G.dis_i)$ | The change in the distribution (local vs global) at interval $i^{th}$ |
| $\Phi^{max}$ | The upper boundary (split) threshold |
| $\Phi^{min}$ | The lower boundary (merge) threshold |
| $size$ | The total number of data tuples (data entries) in a time window |
| $count_i$ | The number of data tuples (entries) falling to the interval $i^{th}$ in a time window |

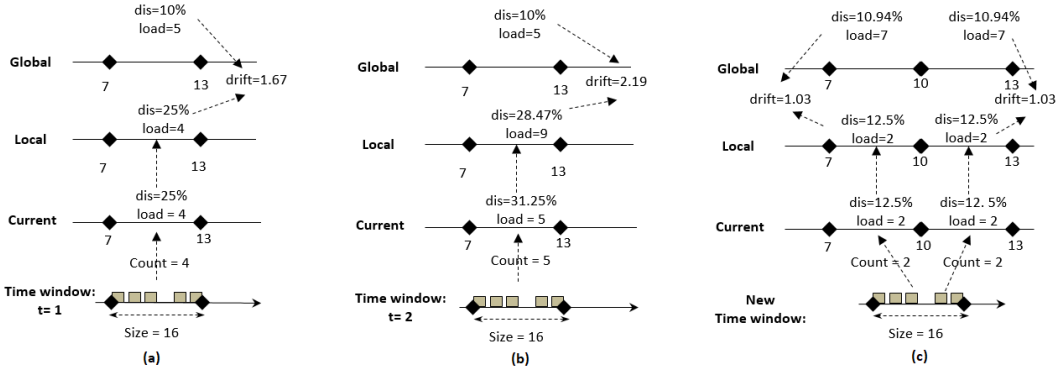


**Figure 4: Online Interval tuning Algorithm applied on the interval $[7 \div 12]$ of LDI shown in Figure 2b.**

The weights for global and local normalized frequency are based on the ratio between the amount of data observed so far. More data represents stronger evidence for $L$ or $D$:

$$w_{L_i} = \frac{L.load_i}{L.load_i + G.load_i} \tag{4}$$

$$w_{G_i} = \frac{G.load_i}{L.load_i + G.load_i} \tag{5}$$

where $L.load_i$, $L.dis_i$, $G.load_d$ and $L.dis_i$ are described in Table 4. This relative distribution drift($i$) is used to determine whether the interval ranges should be modified. While drift remains near 1, no change is needed; but as the value goes above $\Phi^{max}$ or below $\Phi^{min}$, the intervals are split or merged accordingly:

$$Action(i) = \begin{cases} \text{drift}(i) \geq \Phi^{max}, & \text{Split.} \\ \Phi^{max} > \text{drift}(i) > \Phi^{min}, & \text{Skip.} \\ \text{drift}(i) \leq \Phi^{min}, & \text{Merge.} \end{cases} \tag{6}$$

where the action applies to the i$^{th}$ interval. In practice, our current thresholds were set as $\Phi^{max} = 2$ and $\Phi^{min} = 0.3$. In Algorithm 1, after each time window $t$, once the $count_i$ data tuples are accumulated to local $L$, the drift for interval $i$ will be calculated to determine whether a split/merge is needed ($Phase2$ – lines [11–25]). If a split is required at an interval $i$ (lines [12–18]), the current data distribution in local $L$ at this position will be accumulated to the global distribution $G$, then split the current interval. Similarly, if a merge is required at an interval $i$ (lines [20–25]), the current data distribution in local $L$ at this position will also be accumulated to the global distribution $G$ (lines [20–24]), then merge the current interval to its left or its right. Finally, if there is any change in the global $G$, then we propagate this change to the current distribution $C$ (lines [26–27]), to make sure the intervals in $C, L, G$ are same.

Figure 4 shows an example of how distributions are updated and applied to the interval boundary $[7 \div 12]$ (the third interval $i = 3$) of LDI shown in Figure 2b. In this example, we consider three time windows $t = 1$, $t = 2$ and $t = 3$, each time window has 16 data entries. The sample data is shown in Figure 1. The Figure 4(a) shows the behaviors of the algorithm when receiving the first 16 data entries. Since in this first 16 data entries (see Figure 1), there are 4 keys falling to $[7, 13)$, the current normalized frequency $C.dis_3 = 4/16 = 25\%$ and the current count $C.load_3 = 4$. These

**Algorithm 1:** Dynamic-interval tuning process

```
1  Interval-Tuning ():
       input  : Global partitioning G, Local distribution O,
                Current distribution C
       output : New G, O
2      G = {G.b_i, G.dis_i, G.load_i} //Global intervals
3      L = {L.b_i, L.dis_i, L.load_i} //Local intervals
4      C = {C.b_i, C.dis_i, C.load_i} //Current intervals
5      Phase 1: Accumulate the current distribution to local
       intervals
6      foreach ({L.dis_i, L.load_i}, {C.dis_i, C.load_i}) in {L, C} do
7          L.dis_i = (L.dis_i * L.load_i +
8              C.dis_i * C.load_i) / (L.load_i + C.load_i)
9          L.load_i += C.load_i
10     Phase 2: Check the drift of local global, tune the global
       intervals if necessary
11     foreach ({L.dis_i, L.load_i}, {G.dis_i, G.load_i}) in {L, G} do
12         if (drift(i) ≥ Φ^max) then
13             G.dis_i = (G.dis_i * G.load_i +
14                 L.dis_i * L.load_i) / (G.load_i + L.load_i)
15             G.load_i += L.load_i
16             L.load_i = 0
17             L.dis_i = 0
18             Split current interval i
19         else if (drift(i) ≤ Φ^min) then
20             G.dis_i = (G.dis_i * G.load_i +
21                 L.dis_i * L.load_i) / (G.load_i + L.load_i)
22             G.load_i += L.load_i
23             L.load_i = 0
24             L.dis_i = 0
25             Merge the current interval i to the left or to the
               right
26     if there is at least a split or merge in G then
27         C = G
```

current distribution and load will be accumulated to the local $L$. Next, the distribution $drift(L.dis_3, G.dis_3) = 1.67$. Since this $drift$ value does not go above $\Phi^{max}$ or below $\Phi^{min}$, there is no split or merge after this time window.

In the next time window $t = 2$ (see Figure 4), since there are 5 keys falling to $[7, 13)$, the current normalized frequency $C.dis_3 = 5/16 = 31.25\%$ and workload $C.load_3 = 5$ will be accumulated with the local (new values: $L.dis_3 = 28.47\%$ and $L.load_3 = 9$). Since $drift(L.dis_3, G.dis_3) = 2.19$ is higher than $\Phi^{max} = 2$, there will be a split in this interval 3 after this time window. Figure 4(c) shows the new intervals after the split and the new values in those intervals when new data arrives.

## 3.3 LDI Maintenance

LDI writes buckets based on the global distribution, but sometimes may need to perform some in-memory merging to write out buckets. Going back to our example in Figure 2b, step 7, data buffer $(1, 29, 8, 4)$ is full but none of the leaf ranges are full. As a result, LDI has to combine data from two sibling leaves with intervals $[1 \div 2]$ and $[3 \div 6]$ to write a bucket $[1, 4]$. This bucket is now only approximately sorted because it contains a range of data that is larger than a single leaf range. Similarly, in step 12, a non-compacted bucket $[12, 16]$ is created that will have to be rebuilt. In both these situations basically the data that arrived fell into each interval, but there was not enough data to create a compacted bucket. Based on pigeon-hole principle such a situation is non-avoidable even if the global distribution is perfectly learned. In these cases, non-compacted bucket will be recreated and will need to be compacted by a data maintenance process, which reads buckets and writes buckets.

**Merging:** As defined in [35], the compactness of an approximately sorted dataset is defined through the average relative bucket range factor (ARB):

$$ARB = \frac{\sum_{i=1}^{K} \left| Range\left(Bucket_i^{sorted}\right) \right|}{\sum_{i=1}^{K} \left| Range\left(Bucket_i^{LDI}\right) \right|} \quad (7)$$

ARB is the ratio of the sum of the range between minimum and maximum values in a bucket for perfectly sorted data versus the range between minimum and maximum values in the same-sized LDI bucket. Range is defined as the difference between largest and smallest value in a bucket. Data is accessed in bucket units. Thus, if approximately sorted data does not have any values in a wrong bucket ARB will have a perfect score of 1.

LDI performs a sort-merge operation by combining all non-compacted buckets, in order to approximate an improved ARB measure. In the context of LDI, a bucket is considered to be non-compacted if it has at least one data key out of the key range where this bucket is indexed (i.e., the bucket had to be merged with another leaf interval). For instance, as shown in the Figure 2b, buckets [1-4] and [12-16] are non-compacted bucket as the data key 4 and 16 are out off the key ranges (i.e., [1-2] and [7-12]) where those buckets are indexed.

# 4 LDI ARCHITECTURE

Figure 5 give an overview architecture of LDI: (i) Data-clustering component generates buckets with data distribution as discussed in Section 3; (ii) Bucket indexing indexes the buckets using interval B-Tree (Section 4.1; and (iii) I/O handler (Section 4.2). LDI receives data tuples and generates buckets. The created buckets are then written to disk by I/O handler, while buckets metadata are indexed in Bucket indexing.

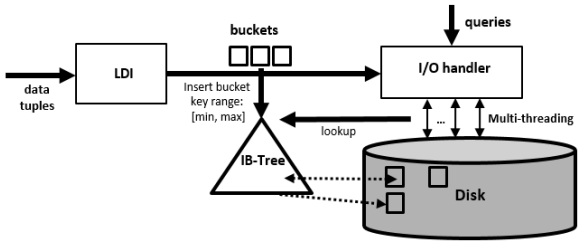


**Figure 5: The overview architecture of LDI.**

## 4.1 Bucket indexing for Query Performance

Buckets generated by LDI are stored on disk and indexed by interval B-Trees [1] (i.e., IB-Trees), in order to further improve the query performance. In LSM typically bloom filters are used to improve query performance. Since analytic workloads are predominantly range queries, an IB tree is more suitable. Incoming buckets are indexed using their key ranges. The range of values $[L, H]$ of a node is the key range of the low key boundary of buckets in its sub-tree. Meanwhile, $Max_i$ of a node indicates the highest value of the high key boundary of buckets in its sub-tree. Using this max value bounds the search.

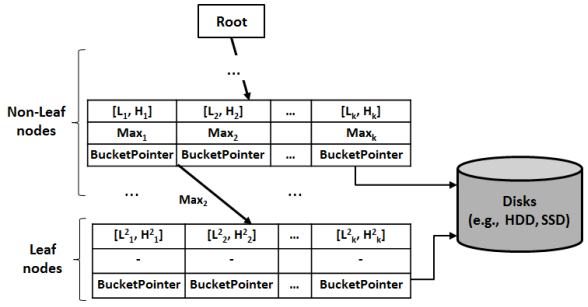


**Figure 6: The data structure of IB-Tree.**

Figure 6 illustrates the adaptation of IB-Tree applied in our context. The size of node is a factor of page size (i.e., $k * 4KB$) so that all nodes in IB-Tree can be efficiently stored on disk. The node's entry in a IB-Tree keeps its data (i.e., $BucketPointer$) [1] in all types of nodes (leaf or non-leaf nodes).

There are a few of advantages of using IB-Tree. First, IB-Tree indexes on the bucket (i.e., group of tuples) instead of projection

[1] $BucketPointer$ points to physical bucket stored on disk (composed of $fileID$ and $offset$).

tuple, its size is reduced by the number of tuples in a bucket. Second, it reduces the time to search for a bucket from $O(N)$ to $O(logN)$ [1, 11] (similar to the search cost in B-Tree). Third, IB-Tree is designed for either full or partial loading in memory.

## 4.2 I/O handler

In most computer system, the storage layer is the major bottleneck; more so with large-scale database systems. While reducing I/O amount improves system performance, the strategy for performing I/O operations also plays crucial role in determining execution time. Current solutions optimize the I/O operations by favoring large granularity and reducing random I/O [25, 30, 34]. The I/O handler is designed to further improve I/O through the following principles:

**Avoid interference between concurrent read and write operations:** Many DBMSes support concurrent execution of insert and read query requests. However, recent experiments show that mixing of parallel reads and writes has a strong negative impact on throughput of storage devices [9]. In our approach, concurrent insertion and read queries are allowed, but there is no mixing of insertion and read queries.

**Concurrent transactions:** Concurrent handling of queries can be exploited to eliminate transfer of data which is redundant among requests. Although multiple identical queries sent to a DBMS at the same time are unlikely, it is common to have overlap in data access by different concurrent requests. To reduce query read cost multiple requests for overlapping data is eliminated by combining requests for the same buckets.

Moreover, requests for different data that are physically stored in close proximity should be grouped, since modern storage devices can efficiently combine such access. To this end, we collect the list of desired buckets and group them into co-located bucket groups (note that bucket is the unit of access, interchangeable with block). Since random I/O access incurs additional overhead, we automatically switch to reading an entire group of buckets (including some extraneous data), based on a threshold $\theta$. The $\theta$ threshold is determined for each specific storage device, using the following formula:

$$\theta = \frac{T_{Read}(Group)}{T_{Read}(Bucket)} \tag{8}$$

where $T_{Read}(Group)$ is the time to read a bucket group and $T_{Read}$ is the time to read a bucket.

**Exploit internal parallelism of storage devices:** We explain how we choose the thread levels in SSDs. SSD has many internal levels of parallelism. Different manufactures or even different devices products may have variations in channels, chips, dies, planes, allocation schemes (how data is allocated on SSD) and so on. Furthermore, as a user, we have no control on how data is physically placed on SSD, which depends on firmware implementation and current usage of SSD.

Nevertheless, choosing the right number of threads can improve the number of planes used, taking advantage of high-bandwidth I/O channel bus and parallelism channel. Given SSD specifications, Formula 9 defines the number of threads used in our model. The idea is to increase the probability of using many planes, while keeping the number of threads low due to the overhead of switching threads.

$$N_{thread} = \frac{N_{channel} * N_{chip} * N_{die} * N_{plane}}{\lambda} \tag{9}$$

Where $N_{channel}$, $N_{chip}$, $N_{die}$ and $N_{plane}$ are the number of channels, chips, dies and planes in an SSD and $\lambda$ is the overhead costs of planning, switching, and buffering temporary results of multiple thread in an SSD-controller.

In theory, CPU and main memory overheads in an SSD controller are considered negligible. Thus, in our evaluation we set $\lambda = 1$. In practice, these costs may have a noticeable effect on performance ($\lambda > 1$). In order to determine a specific value of $\lambda$ for a drive, one would rely on an SSD-specific third party benchmark. For HDDs, we rely on Native Command Queuing (NCQ) for sending parallel requests.

# 5 EVALUATION

*Purpose.* Our experiments 1) demonstrate the significant improvement in data loading performance for *LDI* index, 2) compare read query performance for *LDI* index to current state-of-the-art approaches, and 3) present the benefits of parallelism achieved by our I/O handler in *LDI*.

*Setup.* In our evaluation, we implemented two versions of basic column-oriented DBMS with *LDI* in C++: one without disk-level parallelism (*LDI*) and other with our I/O handler parallelism optimization (*LDI_Par*). We also implemented a basic column-oriented DBMS (*Sorted_Col*), in which all projections are sorted on the indexed key. Queries executed in *Sorted_Col* are expected to have the best possible performance as columns are fully sorted. We also deployed a column-store with LSM-Tree-like indexing (*Col_LSM-like*) that used the same datasets.

*Systems.* Experiments were conducted on a desktop computer with an Intel Core i7-3770 3.4Ghz (8 cores), 8GB of main memory, 1 TB SATA HDD and 256 GB Intel SATA SSD 600p, and Ubuntu 16.04 64-bit operating system.

**Table 5: NYC dataset sizes.**

| Table | Records(M) | Size(GB) |
|---|---|---|
| T_A | 1.5 | 0.3 |
| T_B | 15 | 3 |
| T_C | 30 | 6 |
| T_D | 59 | 11.8 |
| T_E | 148 | 29.6 |

*Dataset & Queries.* Experiments used real-world data from the New York City Taxi (NYC) dataset [31]. Table 5 summarizes the five table sizes used. Three different projections were deployed for each indexing approach: $Projection1$ (29 columns, indexed on $ID$), $Projection2$ (4 columns, indexed on $trip\_distance$) and $Projection3$ (5 columns, indexed on $total\_amount$).

Table 6 summarizes the set of read queries used to measure performance. Key range refers to a delta in the indexed attribute value, i.e., range between $X$ and $X+ < keyrange >$, where $X$ refers to the value of the indexed attribute ($trip\_distance$, $ID$, or $total\_amount$).

Selectivity refers to the ratio between the number of query result tuples to the total database tuples.

**Table 6: Query Range and Selectivity.**

| Range Query | Key Range | Selectivity |
|---|---|---|
| Q1 | 0.003 | 0.003 |
| Q2 | 0.005 | 0.006 |
| Q3 | 0.020 | 0.015 |
| Q4 | 0.030 | 0.026 |
| Q5 | 0.055 | 0.050 |
| Q6 | 0.065 | 0.067 |
| Q7 | 0.075 | 0.083 |
| Q8 | 0.085 | 0.100 |
| Q9 | 0.100 | 0.130 |
| Q10 | 0.110 | 0.150 |

## 5.1 Loading Costs

Figures 7 and 8 summarize the loading costs for the three approaches on both HDD and SSD, respectively. To provide a baseline reference, we also include Write$_{MAX}$, the maximum sequential write speed for each storage device. The *Sorted* runtimes include the load time plus a one-time clustering operation performed at the end of a bulk-load. Similarly, the *Col_LSM-like* runtimes include the costs for reorganizing data in all projections. It's worth emphasizing that the results are for one-time bulk-load ingestion, whereas continuous data ingestion or incremental data loading will likely lead to additional clustering/optimization overheads in both *Sorted* and *Col_LSM-like*. Meanwhile, *LDI* is an on-line process that does not incur any additional overheads for incremental data loading.

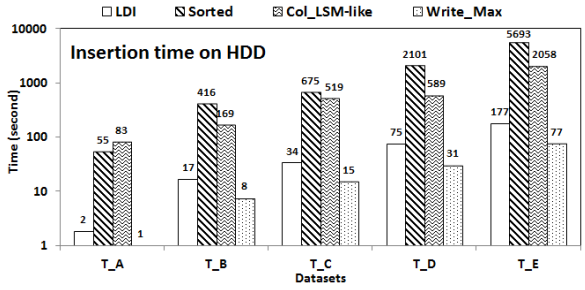


**Figure 7: Loading costs on HDD for different tables.**

***HDD***. We observed that the load times for *LDI* were within a factor of two compared to the disk capacity Write$_{MAX}$. *LDI* significantly improved data load time for the column-oriented DBMS on an HDD; it was, on average, 26 times faster than *Sorted* across all tables (note that the load times are shown on a logarithmic scale). The advantage of *LDI* comes from avoiding data maintenance and clustering overhead during ingestion. *Sorted* includes the loading time and a cost of clustering at the end of bulk-loading. Similarly, loading runtimes in *Col_LSM-like* include merge costs. Moreover, incremental loading of data (versus a one-time bulk load) can only degrade the performance of both *Sorted* and *Col_LSM-like*. In contrast, *LDI* load cost is always based on the amount of data it ingests.

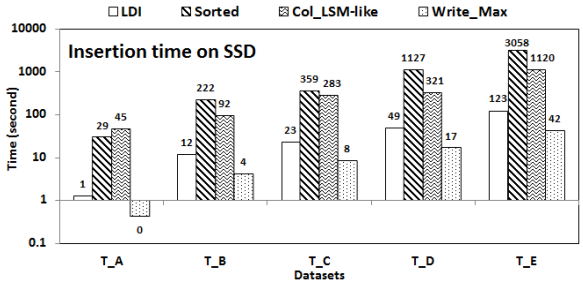


**Figure 8: Loading runtimes on SSD with different tables.**

***SSD***. All runtimes are significantly improved on SSD (see Figure 8) because the throughput of SSD is much higher than throughput of HDD. As with HDD, *LDI* significantly outperforms *Sorted* and *Col_LSM-like*. The advantage of *LDI* in SSD load times is similar to its advantages in the HDD evaluation, because the overheads of *Sorted* and *Col_LSM-like* approaches are caused by mostly sequential extra I/O operations.

## 5.2 Read Query Performance

*LDI* eliminates clustering reorganization (sort and merge) costs associated with data loading by creating approximately clustered buckets. Since *LDI* reads queried data in bucket units, it may read extraneous tuples stored in the buckets that contain some of the data in queried range. As a result, *LDI* queries may incur a small penalty in comparison to approaches that strictly cluster data. In this evaluation, we compare *LDI* to the best possible query performance of *Sorted*. We ran queries against two different projections: *Projection*1 (29 columns) and *Projection*2 (4 columns) with all competitors: *LDI*, *LDI_par* (with parallelism optimization), *Sorted* and *Col_LSM-like*. Query runtimes for both HDD and SSD are shown in Figure 9.

The first observation is that *Sorted* typically has the fastest query runtimes in queries on both HDD (i.e., Figure 9(a) and 9(c)) and SSD (i.e., Figure 9(b) and 9(d)). This is because data is perfectly sorted in *Sorted*. *LDI_par* is only slightly slower than than *Sorted* on smaller queries and on HDD (see Figure 9 (a) and (c)). However, *LDI_par* actually outperforms *Sorted* with larger queries, especially on SSD (see Figure 9 (b) and (d)). *LDI_par* can outperform Sorted by leveraging our parallelism optimization, which is most effective on larger queries (in terms of amount of data) and on SSD.

As expected, the query runtimes of *LDI* (without parallelism) are slightly slower than those of *Sorted* because: 1) *LDI* data is approximately sorted in buckets, causing it to read some extraneous data that is not required by the query, and 2) read operations are done in bucket granularity instead of sequential reads in *Sorted*. However, on SSD, when the performance of random I/O and sequential I/O are almost similar, *LDI* performance is better than *Sorted* with large queries (see Figure 9 (c) and (d)). This comes from our I/O handler in *LDI* that takes advantage of internal parallel mechanism on SSD.

*Col_LSM-like* shows the slowest query performance among evaluated approaches, for all tested queries and on both HDD and SSD.

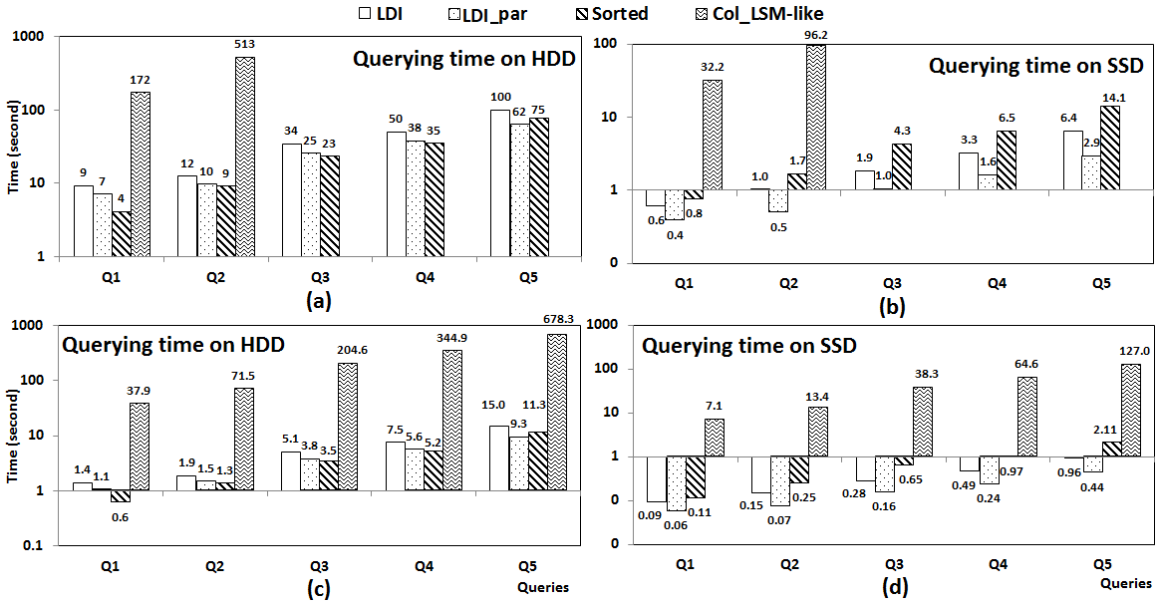


**Figure 9: Runtimes of different queries on HDD and SSD at different projections: (a) Runtimes on *Projection*1 with HDD; (b) Runtimes on *Projection*1 with SSD; (c) Runtimes on *Projection*2 with HDD; and (d) Runtimes on *Projection*2 with SSD.**

It takes long time to query *Projection*1 (see Figure 9(a) and 9(b)). For better readability we only show the runtimes of queries $Q1$ and $Q2$, as other queries take even longer. *Col_LSM-like* runtimes on *Projection*2 are faster (see Figure 9(c) and 9(d)), but still slower than other methods.

## 5.3 Effectiveness of concurrence and parallel processing

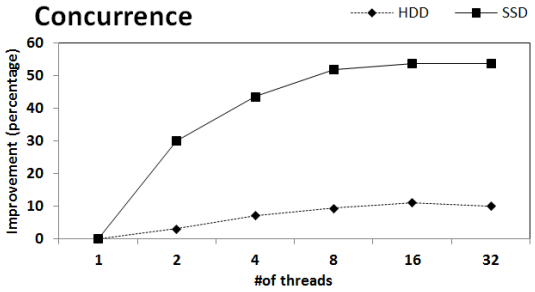


**Figure 10: Runtime improvements on HDD and SSD.**

As shown in Section 5.1, parallelism has a strong impact on ingestion performance. The experiments in this section aim to evaluate the effectiveness of concurrence and parallel processing in read queries. In order to show how *LDI* behaves in presence of concurrence requests, by de-duplicating overlapping data and by leveraging concurrency, we created a set of 40 queries with selectivity ranging from 0.2% to 5%. We then run these queries, varying the number of threads from 1 to 32. Figure 10 presents the runtime improvements (percentage improvement on y-axis) for both HDD and SSD as the number of threads increases (x-axis).

As Figure 10 shows, while improvement increases (the runtime decreases) as the number of threads increases, the best number of threads in HDD is 8 or 16 threads (improvement of 9% or 11%), while on SSD the best number of threads is 16 or 32 (improvement of 53%). The results confirm that 1) concurrent processing improves query performance both in HDD and in SSD, and 2) SSD has a greater capability for exploiting concurrence compared to HDD.

We also observe that increasing the number of threads eventually becomes counterproductive. For example, in our experiments 32 threads on HDD and 64 threads on SSD lead to query deterioration. In fact, the effectiveness of concurrence comes from parallelism and de-duplication. However when as the number of threads increases, the overheads of multi-threading will negate the benefits of concurrence. Furthermore, a large number of threads may require a significant amount of memory to store temporary query results.

In order to examine the impact of the number of threads in reading/writing from/to storage devices, we vary the number of threads that are allowed to read/write at the same time using queries from Table 6. Figures 11 and 12 show the runtimes of different queries as we vary the number of parallel threads (x-axis shows the number of threads, y-axis shows the execution time in seconds).

Figure 11 presents the results for Q1 through Q5 (queries described in Table 6), while Figure 12 shows the results for Q6 through Q10. Both HDD and SSD work better with multiple threads. The best number of threads on HDD and on SSD is 8 and 16 threads, respectively. Continuing to increase the number of threads offers very little benefit or potentially begins to slow queries down due to the overhead costs of multi-threading.

Additionally, the results in Figures 11 and 12 further confirms that SSD has better support for parallelism as compared to HDD, since the benefits of parallelism are larger on SSD vs HDD. In particular, the improvement of parallelism optimization on SSD

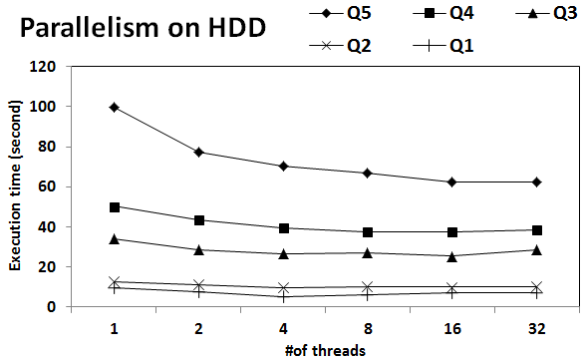


**Figure 11: Parallel on HDD.**

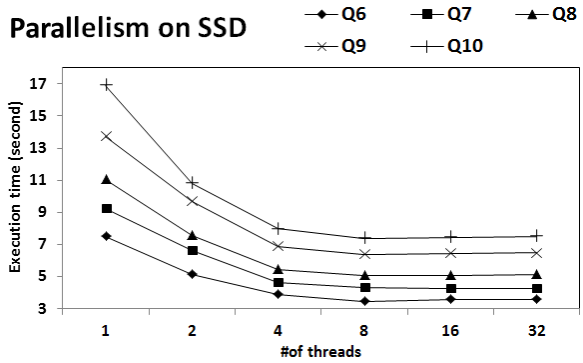


**Figure 12: Parallel on SSD.**

is around 54% (at 16 threads), while on HDD the improvement is around 25% (at 8 threads).

# 6 RELATED WORK

In this section, we first present some prominent columnar stores in Section 6.1, then summarize some work about log-structure merge tree, i.e. the most wide-applied write-optimized indexing methods, in Section 6.2. Last, we discuss some related work about data distribution in Section 6.3.

## 6.1 Columnar Stores

The ideas of column-oriented hierarchical have been widely applied in both academic [19, 33] and commercial [23, 24, 29, 37]. MonetDB [19] is an in-memory columnar relational DBMS that leverages the large main memories of modern computer systems in query processing, while the database is persistently stored on disk, showing outstanding performance. However, similar to other in-memory column-oriented DBMSes such as SAP HANA [16], Peloton [3], DB2 BLU [29] or Microsoft SQL Server Column stores [24], its application is limited due to many reasons: (i) require huge amount of main memory; and (ii) the volatility of main memory. Recently, in the same context as MonetDB, Vectorwise [37] was proposed, aiming at vectorizing execution to operate on vector of data instead of separate tuples and further improving in storage model. However, it has some drawbacks like MonetDB, as it's in-memory columnar store. Furthermore, most of these approaches are designed for query-mostly workload. Continuous insertion workflows often lead to dramatic degradation in performance.

C-Store [33] and its recent commercial extension (i.e., Vertica [23]) are on-disk columnar databases that apply different sort orders on each projections (i.e., groups of columns), different compression method for each column in order to improve the compression ratio, fully support aggregation operations on compressed data. However, similar to other column stores, incremental data loading may require heavy data clustering/re-organization processes. Even though, these processes can be run as background processes in Vertica, it negatively impacts DBMS performance.

## 6.2 Log Structured Merge Trees

While many DBMSes including columnar stores suffer from poor write performance, the log-structure merge tree (LSM-Tree) [27, 32] is a common solution for this problem. The main idea is to transform small random writes into large sequential writes by buffering and reorganizing data in a large buffer before flushing them to disk at a batch. Data on disk is structured as many levels, and data will be merged to the higher levels as the data grows. Compared to clustered index (or primary index) requiring clustering process every time a new data is inserted, LSM-Tree accumulates loading data in the buffer (level 0) before merging it to higher levels on disk. This delaying technique significantly reduce the re-organization cost as in clustered index. Moreover, all operations in LSM-tree are done in batch (read/write with large I/O), making its performance much better in any type of storage environment (e.g., HDD or SSD).

Due to those reasons, LSM-Tree has been widely applied not only in columnar DBMSes, but also in other types of DBMSes such as relational DBMSes, Columnar stores or Key-Value stores. For example, LevelDB [18], BigTable [8], HBase [2], Cassandra [22] are some key-value stores apply LSM-Tree, while MySQL [28] and SQLite4 [12] are relational DBMSes that support LSM-Tree indexing. Other enhanced variations of LSM-Tree used in Monkey [13] and Dostoevsky [14] to further improve the DBMS performance by using BloomFilter and changing merging policies. However, the main drawback of LSM-Tree, i.e., large write amplification [2] still remains.

Similarly in TRIAD [5], couple of improvement techniques have been applied to reduce the write amplification such as (i) keeping the hot-entries longer in the main memory; (ii) changing the tired merging policy by considering the overlapping between runs in a level; and (iii) optimize the write in commit-log. Unfortunately, the gain comes with a price. Lower write amplification is archived by using a variation of tiered policy (i.e., favor the write performance) and by scarifying the look up performance (the higher the merge overlapping threshold, the more degration in look up). Furthermore, similar to Dostoevsky [14], this method only reduces the high write amplification of LSM-Tree, but cannot completely avoid this issue.

[2] Write amplification is the ratio of total write IO performed by the DBMS to the total data in the DBMS. High write amplification increases the loading cost on storage devices.

### 6.3 Data distribution

Knowing data distributions is crucial important in database systems and data streaming. However, accurately record data distributions is expensive, leading to many approaches for approximately capture data distribution (called incremental histograms) [15, 17, 26]. For example, Gibbons et al. [17] proposed an approximate histograms maintained in the present of data insertion and a merge and split technique for adjusting histogram buckets according to the data insertion. Meanwhile, Mousavi et al. [26] introduced an approximate approach for incrementally approximate compute equi-depth histograms over sliding windows.

Those histogram methods have been applied in many aspects of database systems such as selectivity estimation (query optimization), approximate query answering, join query execution. However, as mentioned in [6, 20], traditional indexing using $B^+$-Tree is not suitable to serve as non-equi-depth histograms. Particularly, straightforwardly applying basic index trees to serve non-equi-depth histograms should significantly degrade indexing performance due to its unbalanced structure or low node occupancy [6]. To the best of our knowledge, there is no work applying histogram information in database indexing.

## 7 CONCLUSION

In this paper, we presented a learned distribution index (LDI) structure as an alternative for LSM tree structure, which is advocated for ingesting data rapidly into projections of a colmumnar database to improve insertion performance. Learning distributions, as we show, avoids the expensive sort-merge phase of LSM. Data ingestion performance, therefore, increase more than order of magnitude in comparison with other methods while query performance is as comparable as LSM-like indexing approaches. The cost of learning distribution may lead to some non-compacted buckets, causing the LDI maintenance. However, this cost of maintenance is reasonable as only the non-compacted buckets are need to be inverted.

## REFERENCES


[1] Chuan-Heng Ang and Kok-Phuang Tan. 1995. The Interval B-tree. *Inf. Process. Lett.* 53, 2 (Jan. 1995), 85–89. https://doi.org/10.1016/0020-0190(94)00176-Y

[2] Apache. [n. d.]. HBase. https://github.com/google/leveldb. ([n. d.]). [Online; accessed 25-Mar-2019].

[3] Joy Arulraj, Andrew Pavlo, and Prashanth Menon. 2016. Bridging the Archipelago Between Row-Stores and Column-Stores for Hybrid Workloads. In *SIGMOD '16*. ACM, New York, NY, USA, 583–598. https://doi.org/10.1145/2882903.2915231

[4] Oana Balmau, Diego Didona, Rachid Guerraoui, Willy Zwaenepoel, Huapeng Yuan, Aashray Arora, Karan Gupta, and Pavan Konka. 2017. TRIAD: Creating Synergies Between Memory, Disk and Log in Log Structured Key-Value Stores. In *2017 USENIX Annual Technical Conference (USENIX ATC 17)*. USENIX Association, Santa Clara, CA, 363–375. https://www.usenix.org/conference/atc17/technical-sessions/presentation/balmau

[5] Oana Balmau, Diego Didona, Rachid Guerraoui, Willy Zwaenepoel, Huapeng Yuan, Aashray Arora, Karan Gupta, and Pavan Konka. 2017. TRIAD: Creating Synergies Between Memory, Disk and Log in Log Structured Key-Value Stores. In *2017 USENIX Annual Technical Conference (USENIX ATC 17)*. USENIX Association, Santa Clara, CA, 363–375. https://www.usenix.org/conference/atc17/technical-sessions/presentation/balmau

[6] Daniel Barbara, William duMouchel, Christos Faloutsos, Peter J. Haas, Joseph M. Hellerstein, Yannis Ioannidis, H. V. Jagadish, Theodore Johnson, Raymond Ng, Viswanath Poosala, Kenneth A. Ross, and Kenneth C. Sevcik. 1997. The New Jersey Data Reduction Report. *IEEE DATA ENG. BULL* (1997), 3–45.

[7] Edward Bortnikov, Anastasia Braginsky, Eshcar Hillel, Idit Keidar, and Gali Sheffi. 2018. Accordion: Better Memory Organization for LSM Key-value Stores. *Proc. VLDB Endow.* 11, 12 (Aug. 2018), 1863–1875. https://doi.org/10.14778/3229863.3229873

[8] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. 2006. Bigtable: A Distributed Storage System for Structured Data. In *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation - Volume 7 (OSDI '06)*. USENIX Association, Berkeley, CA, USA, 15–15.

[9] F. Chen, R. Lee, and X. Zhang. 2011. Essential roles of exploiting internal parallelism of flash memory based solid state drives in high-speed data processing. In *2011 IEEE 17th International Symposium on High Performance Computer Architecture*. 266–277. https://doi.org/10.1109/HPCA.2011.5749735

[10] Columnar Database. 2019. Columnar Database: A Smart choice for data warehouses. https://www.columnardatabase.com/. (2019). [Online; accessed 25-Mar-2019].

[11] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. *Introduction to Algorithms, Third Edition* (3rd ed.). The MIT Press.

[12] D. Richard Hipp. [n. d.]. SQLite4. https://sqlite.org/src4/doc/trunk/www/index.wiki. ([n. d.]). [Online; accessed 25-Mar-2019].

[13] Niv Dayan, Manos Athanassoulis, and Stratos Idreos. 2017. Monkey: Optimal Navigable Key-Value Store. In *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17)*. ACM, New York, NY, USA, 79–94. https://doi.org/10.1145/3035918.3064054

[14] Niv Dayan and Stratos Idreos. 2018. Dostoevsky: Better Space-Time Trade-Offs for LSM-Tree Based Key-Value Stores via Adaptive Removal of Superfluous Merging. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18)*. ACM, New York, NY, USA, 505–520. https://doi.org/10.1145/3183713.3196927

[15] D. Donjerkovic, Y. Ioannidis, and R. Ramakrishnan. 2000. Dynamic Histograms: Capturing Evolving Data Sets. In *Proceedings of 16th International Conference on Data Engineering (Cat. No.00CB37073)*. 86–. https://doi.org/10.1109/ICDE.2000.839394

[16] Franz Färber, Sang Kyun Cha, Jürgen Primsch, Christof Bornhövd, Stefan Sigg, and Wolfgang Lehner. 2012. SAP HANA Database: Data Management for Modern Business Applications. *SIGMOD Rec.* 40, 4 (Jan. 2012), 45–51. https://doi.org/10.1145/2094114.2094126

[17] Phillip B. Gibbons, Yossi Matias, and Viswanath Poosala. 2002. Fast Incremental Maintenance of Approximate Histograms. *ACM Trans. Database Syst.* 27, 3 (Sept. 2002), 261–298. https://doi.org/10.1145/581751.581753

[18] Google. [n. d.]. LevelDB. https://github.com/google/leveldb. ([n. d.]). [Online; accessed 25-Mar-2019].

[19] Stratos Idreos, Fabian Groffen, Niels Nes, Stefan Manegold, K. Sjoerd Mullender, and Martin L. Kersten. 2012. MonetDB: Two Decades of Research in Column-oriented Database Architectures. *IEEE Data Engineering Bulletin* 35, 1 (2012), 40–45.

[20] Yannis Ioannidis. 2003. The History of Histograms (Abridged). In *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29 (VLDB '03)*. VLDB Endowment, 19–30. http://dl.acm.org/citation.cfm?id=1315451.1315455

[21] Bradley C. Kuszmaul. 014. A Comparison of Fractal Trees to Log-Structured Merge (LSM) Trees. *White Paper* (014).

[22] Avinash Lakshman and Prashant Malik. 2010. Cassandra: A Decentralized Structured Storage System. *SIGOPS Oper. Syst. Rev.* 44, 2 (April 2010), 35–40. https://doi.org/10.1145/1773912.1773922

[23] Andrew Lamb, Matt Fuller, Ramakrishna Varadarajan, Nga Tran, Ben Vandiver, Lyric Doshi, and Chuck Bear. 2012. The Vertica Analytic Database: C-store 7 Years Later. *Proc. VLDB Endow.* 5, 12 (Aug. 2012), 1790–1801. https://doi.org/10.14778/2367502.2367518

[24] Paul Larson, Cipri Clinciu, Campbell Fraser, Eric N. Hanson, Mostafa Mokhtar, Michal Nowakiewicz, Vassilis Papadimos, Susan L. Price, Srikumar Rangarajan, Remus Rusanu, and Mayukh Saubhasik. 2013. Enhancements to SQL Server Column Stores. In *SIGMOD '13*.

[25] Yinan Li, Bingsheng He, Robin Jun Yang, Qiong Luo, and Ke Yi. 2010. Tree Indexing on Solid State Drives. *Proc. VLDB Endow.* 3, 1-2 (Sept. 2010), 1195–1206. https://doi.org/10.14778/1920841.1920990

[26] Hamid Mousavi and Carlo Zaniolo. 2011. Fast and Accurate Computation of Equi-depth Histograms over Data Streams. In *Proceedings of the 14th International Conference on Extending Database Technology (EDBT/ICDT '11)*. ACM, New York, NY, USA, 69–80. https://doi.org/10.1145/1951365.1951376

[27] Patrick O'Neil, Edward Cheng, Dieter Gawlick, and Elizabeth O'Neil. 1996. The Log-structured Merge-tree (LSM-tree). *Acta Inf.* 33, 4 (June 1996), 351–385. https://doi.org/10.1007/s002360050048

[28] Oracle. [n. d.]. MySQL. https://github.com/mysql/mysql-server. ([n. d.]). [Online; accessed 25-Mar-2019].

[29] Vijayshankar Raman, Gopi Attaluri, Ronald Barber, Naresh Chainani, David Kalmuk, Vincent KulandaiSamy, Jens Leenstra, Sam Lightstone, Shaorong Liu, Guy M. Lohman, Tim Malkemus, Rene Mueller, Ippokratis Pandis, Berni Schiefer, David Sharpe, Richard Sidle, Adam Storm, and Liping Zhang. 2013. DB2 with BLU Acceleration: So Much More Than Just a Column Store. *Proc. VLDB Endow.* 6, 11 (Aug. 2013), 1080–1091. https://doi.org/10.14778/2536222.2536233

[30] Iulian Sandu Popa, Karine Zeitouni, Vincent Oria, Dominique Barth, and Sandrine Vial. 2011. Indexing In-network Trajectory Flows. *The VLDB Journal* 20, 5 (Oct. 2011), 643–669. https://doi.org/10.1007/s00778-011-0236-8

[31] Todd W. Schneider. 2016. Unified New York City Taxi and Uber data. https://github.com/toddwschneider/nyc-taxi-data. (2016). [Online; accessed 18-Aug-2017].
[32] Russell Sears and Raghu Ramakrishnan. 2012. bLSM: A General Purpose Log Structured Merge Tree. In *SIGMOD '12*. ACM, New York, NY, USA, 217–228. https://doi.org/10.1145/2213836.2213862
[33] Mike Stonebraker, Daniel J. Abadi, Adam Batkin, Xuedong Chen, Mitch Cherniack, Miguel Ferreira, Edmond Lau, Amerson Lin, Sam Madden, Elizabeth O'Neil, Pat O'Neil, Alex Rasin, Nga Tran, and Stan Zdonik. 2005. C-store: A Column-oriented DBMS. In *VLDB'05*. 12.
[34] Dai Hai Ton-That, Iulian Sandu-Popa, and Karine Zeitouni. 2015. TRIFL: A Generic Trajectory Index for Flash Storage. *ACM Trans. Spatial Algorithms Syst.* 1, 2, Article 6 (July 2015), 44 pages. https://doi.org/10.1145/2786758
[35] Dai Hai Ton That, James Wagner, Alexander Rasin, and Tanu Malik. 2018. PLI$^+$: Efficient Clustering of Cloud Databases. *Distributed and Parallel Databases* (2018). in the third round of DAPD.
[36] Wikipedia. 2019. Business Intelligence. https://en.wikipedia.org/wiki/Business_intelligence. (2019). [Online; accessed 25-June-2019].
[37] Marcin Zukowski and Peter A. Boncz. 2012. Vectorwise: Beyond Column Stores. *IEEE Data Eng. Bull.* 35 (2012), 21–27.

# 8 APPENDIX

## 8.1 Example: Tiered LSM-Tree

The Figure 13 shows the Tiered LSM-Tree behavior during the data ingestion (Sample data D - Figure 1). As shown, there are only two merges at the steps 2 and 4.

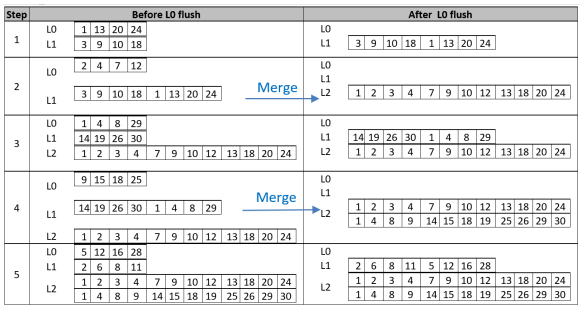


**Figure 13: Data loading with tired and tiered LSM-Trees.**

## 8.2 LSM-Tree merge costs

The the number of I/Os (write and read) of Tiered-LSM and Leveled-LSM are presented in Formulas ( 10 and 12) and ( 11 and 13).

$$n_w^R = \sum_{i=1}^{L-1} \left( \left\lfloor \frac{M}{T^{i-1}} \right\rfloor - \left\lfloor \frac{M}{T^i} \right\rfloor \right) \frac{BT^{i-1}}{P} \tag{10}$$

Proof. Formula 10 can be proved as follows. The Tier-LSM has L levels ($i.e.$, $[0; L-1]$). Since Level 0 is the head of LSM resided main-memory we only consider levels in $[1; L-1]$. Given $M = \lfloor N/B \rfloor$, at level $L = 1$, we have $M$ times level $L = 0$ get full and flush a run to level $L = 1$. Among these $M$ times, there are $\lfloor \frac{M}{T} \rfloor$ times level $L = 1$ get full and all data will be written to level $L = 2$ instead of level $L = 1$. In level $L = 1$, $run = B$. Therefore, we have: $run*(M - \lfloor \frac{M}{T} \rfloor) = \frac{B}{P} * (M - \lfloor \frac{M}{T} \rfloor)$ write IO. At level $L = 2$: we have $\lfloor \frac{M}{T} \rfloor$ times level $L = 1$ get full and flush a run to level $L = 2$. Among these $\lfloor \frac{M}{T} \rfloor$ times, there are $\lfloor \frac{M}{T^2} \rfloor$ times level $L = 2$ get full and all data will be written to level $L = 3$ instead of level $L = 2$. In level $L = 2$, $run = BT$. Therefore, we have: $run*(\lfloor \frac{M}{T} \rfloor - \lfloor \frac{M}{T^2} \rfloor) = \frac{BT}{P} * (\lfloor \frac{M}{T} \rfloor - \lfloor \frac{M}{T^2} \rfloor)$ write IO. At Level $L = i$ ($1 \le i \le L-1$): we have $\lfloor \frac{M}{T^{i-1}} \rfloor$ times level $L = i-1$ get full and flush a run to level $L = i$. Among these $\lfloor \frac{M}{T^{i-1}} \rfloor$ times, there are $\lfloor \frac{M}{T^i} \rfloor$ times level $L = i$ get full and all data will be written to level $L = i+1$ instead of level $L = i$. In level $L = i$, $run = BT^{i-1}$. Therefore, we have: $run*(\lfloor \frac{M}{T^{i-1}} \rfloor - \lfloor \frac{M}{T^i} \rfloor) = \frac{BT^{i-1}}{P} * (\lfloor \frac{M}{T^{i-1}} \rfloor - \lfloor \frac{M}{T^i} \rfloor)$ write IO. Therefore the total number of writes:
$n_w^R = \sum_{i=1}^{L-1} (\lfloor \frac{M}{T^{i-1}} \rfloor - \lfloor \frac{M}{T^i} \rfloor) * \frac{BT^{i-1}}{P}$. □

$$n_w^V = \sum_{i=1}^{L-1} \left\lfloor \left\lfloor \frac{M}{T^i} \right\rfloor \frac{(T-1)T^iB}{2P} + \left( \left\lfloor \frac{M}{T^{i-1}} \right\rfloor \%T \right) \left( \left\lfloor \frac{M}{T^{i-1}} \right\rfloor \%T + 1 \right) \frac{T^{i-1}B}{2P} \right\rfloor \tag{11}$$

Proof. Formula 11 can be proved as follows. The Leveled-LSM has L levels ($i.e.$, $[0; L-1]$). Since Level 0 is the head of LSM resided main-memory we only consider levels in $[1; L-1]$. Given $M = \lfloor N/B \rfloor$, at level $L = 1$, we have $A_1 = \lfloor \frac{M}{T} \rfloor$ times level $L = 1$ get full and $B_1 = M\%T$ runs are still remaining in level $L = 1$. Before each time level $L = 1$ get full, it needs to received $T$ runs from level $L = 0$. Each time level $L = 1$ receives a run from level $L = 0$, it reads exiting data and re-write everything again to this level except when it's full. This means before each time $L = 1$, get full. It needs to write $\frac{run}{P} * [(1+2+...+(T-1)]$. Meanwhile, in level $L = 1$, $run = B$. The number of writes in $A_1$ is $\lfloor \frac{M}{T} \rfloor * \frac{B}{P} * \sum_{t=1}^{T-1} t = \lfloor \frac{M}{T} \rfloor * \frac{B(T-1)T}{2P}$. Similarly, the cost of writing $B_1$ runs in level $L = 1$ is $\frac{run}{P} * [1+2+...+B_1] = \frac{B}{P} * [1+2+...+M\%T] = \frac{B}{P} * \frac{M\%T*(M\%T+1)}{2}$. Therefore, the total number of writes in level $L = 1$: $\lfloor \frac{M}{T} \rfloor * \frac{B(T-1)T}{2P} + \frac{M\%T*(M\%T+1)B}{2P}$. At Level $L = 2$: Similarly, we have $A_2 = \lfloor \frac{M}{T^2} \rfloor$ times level $L = 2$ get full and $B_2 = \lfloor \frac{M}{T} \rfloor \%T$ runs are still remaining in level $L = 2$. In level $L = 2$, $run = B * T$. So the total number of writes in level $L = 2$: $\lfloor \frac{M}{T^2} \rfloor * \frac{B(T-1)T^2}{2P} + \lfloor \frac{M}{T} \rfloor \%T * (\lfloor \frac{M}{T} \rfloor \%T + 1) \frac{BT}{2P}$. At Level $L = i$ ($1 \le i \le L-1$): We also have $A_i = \lfloor \frac{M}{T^i} \rfloor$ times level $L = i$ get full and $B_i = \lfloor \frac{M}{T^{i-1}} \rfloor \%T$ runs are still remaining in level $L = i$. In level $L = i$, $run = B * T^{i-1}$. So the total number of writes in level $L = i$: $\lfloor \frac{M}{T^i} \rfloor * \frac{B(T-1)T^i}{2P} + \lfloor \frac{M}{T^{i-1}} \rfloor \%T * (\lfloor \frac{M}{T^{i-1}} \rfloor \%T + 1) \frac{BT^{i-1}}{2P}$
Therefore the total number of writes:
$n_w^V = \sum_{i=1}^{L-1} \lfloor \frac{M}{T^i} \rfloor * \frac{B(T-1)T^i}{2P} + \lfloor \frac{M}{T^{i-1}} \rfloor \%T * (\lfloor \frac{M}{T^{i-1}} \rfloor \%T + 1) \frac{BT^{i-1}}{2P}$. □

$$n_r^R = \sum_{i=1}^{L-1} \left\lfloor \left\lfloor \frac{M}{T^i} \right\rfloor \frac{(T-1)B}{P} \right\rfloor = m^T \left\lfloor \frac{(T-1)B}{P} \right\rfloor \tag{12}$$

$$n_r^V = \sum_{i=1}^{L-1} \left\lfloor \left\lfloor \frac{M}{T_i} \right\rfloor \frac{(T-1)T^iB}{2P} + \left( \left\lfloor \frac{M}{T^{i-1}} \right\rfloor \%T \right) \left( \left\lfloor \frac{M}{T^{i-1}} \right\rfloor \%T - 1 \right) \frac{T^{i-1}B}{2P} \right\rfloor \tag{13}$$

Similar methods can be applied to have the total number of IOs (read) for Tiered-LSM and Leveled-LSM shown in Formulas 12 and 13.