SortComp (Sort-and-Compress)
-- Towards a Universal Lossless Compression Scheme for Matrix and Tabular Data

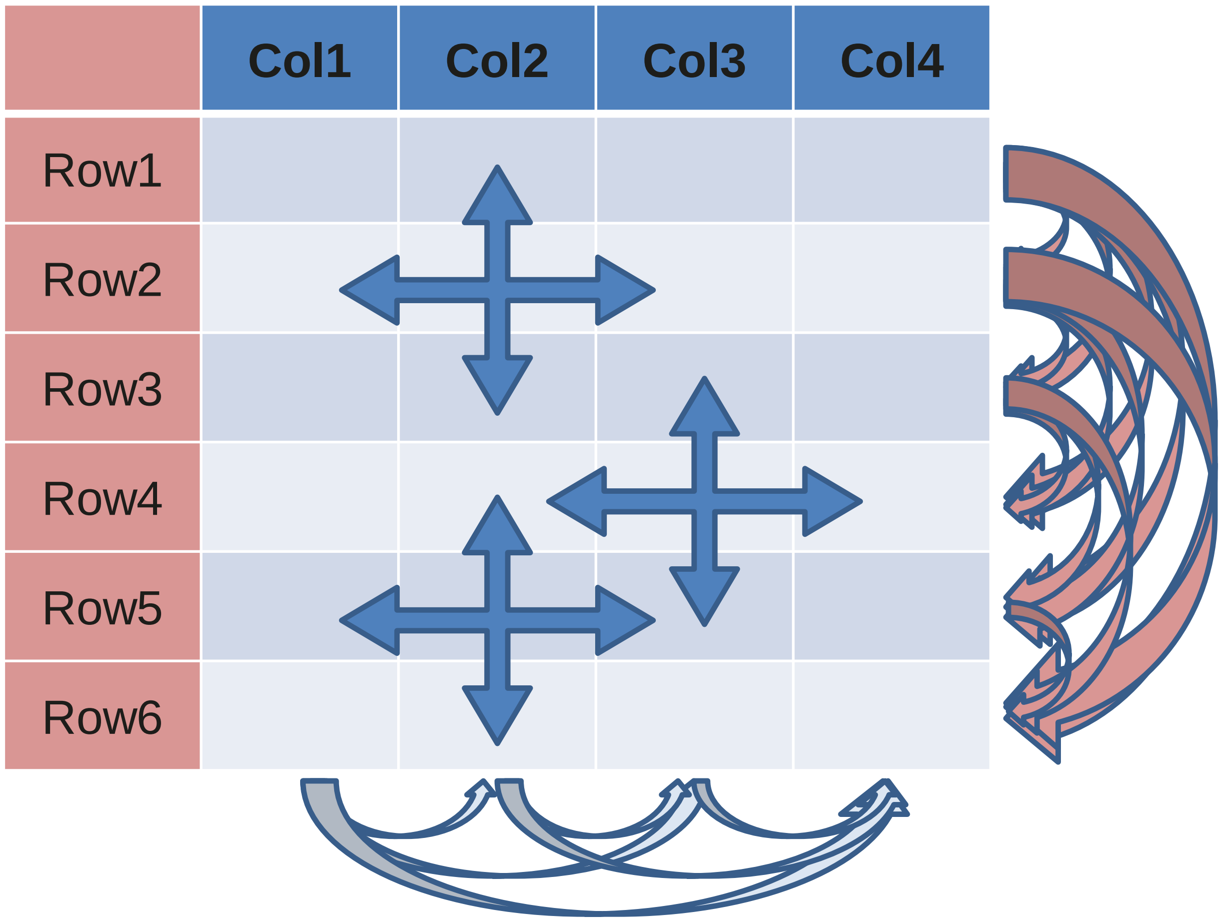
Xizhe CHENG, Sian-Jheng LIN, Jie SUN
cheng.xizhe@huawei.com

Theory Lab.,
Hong Kong R&D Center,
Huawei Technologies Co. Ltd.

HUAWEI

To Compress Tabular Data

2D Tabular Data Storage



- **A Great Amount of Data can be 2-Dimensional**

  Spread sheets, pictures, relational database files …

- **Tabular Data Characteristics**

  Data dependencies along both row-wise and column-wise directions;

- **Posing Challenges to General-Purpose Data Compression Algorithm**

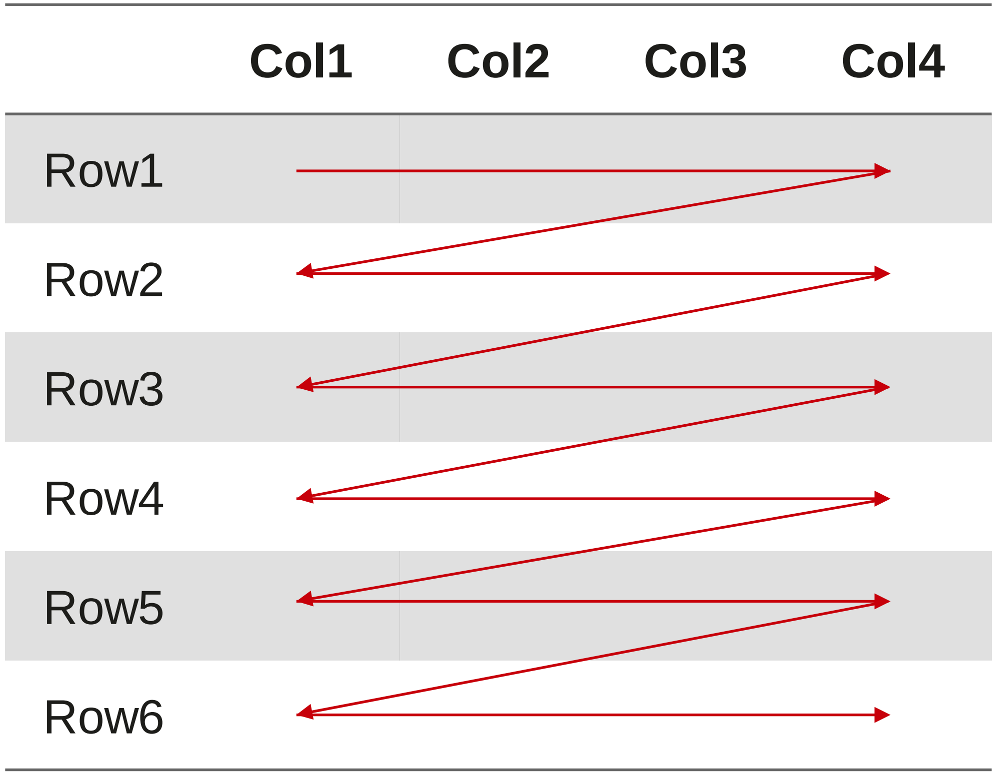
  General-purpose data compression algorithms assume 1D nature of the datastream, losing track of the column-wise data dependencies;

HUAWEI

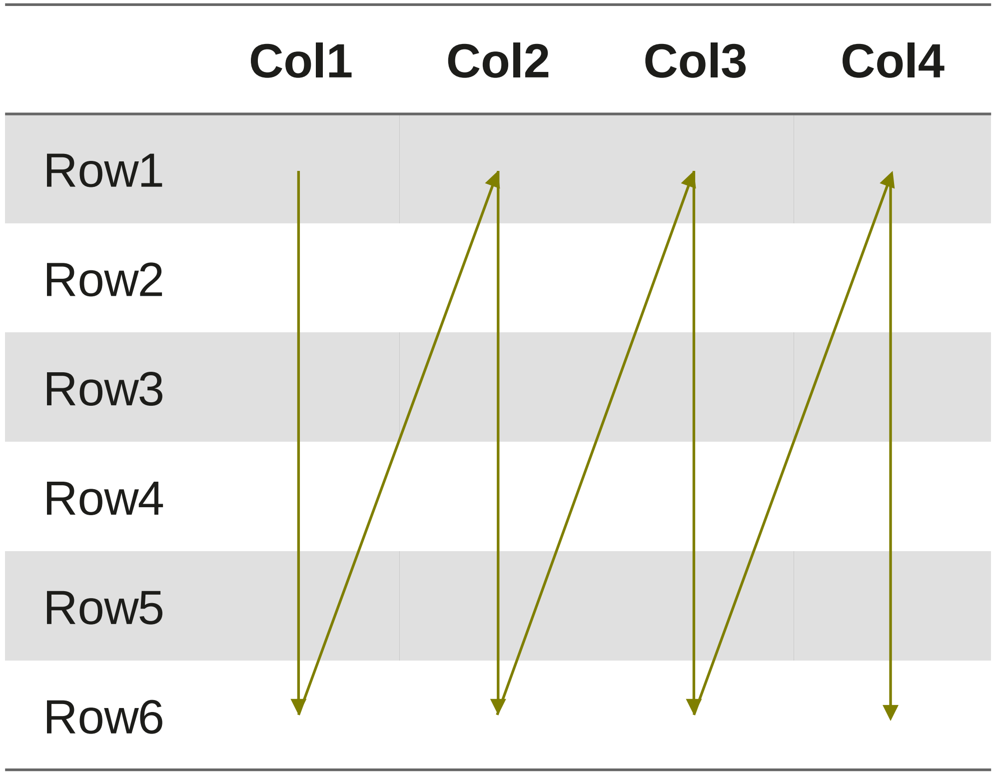
Data Locality Matters

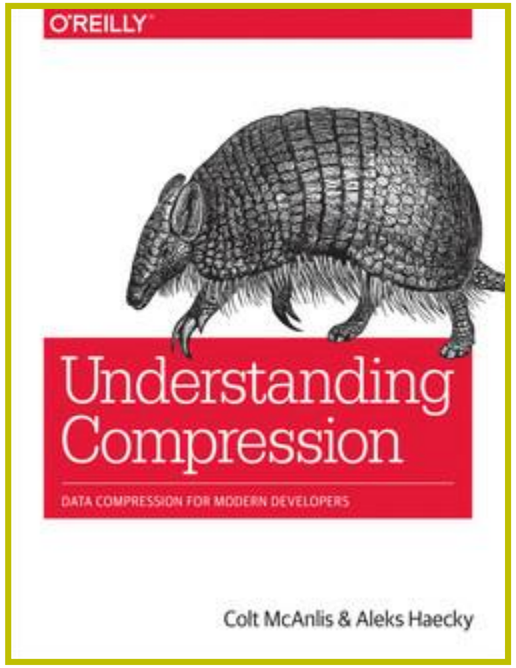
This leads us to a very important place in the compression world, the concept that locality matters.[4] As data is created in a linear fashion, there's a high probability that

Row-Major Traversing



Column-Major Traversing





Which one to choose?

- Different traversing strategies suits different data localities the best

HUAWEI

Related Works

- Table Compressors Based on Row-Reordering

  Daniel Lemire, Owen Kaser, and Eduardo Gutarra, "Reordering rows for better compression: Beyond the lexicographic order," *ACM Trans. Database Syst.*, vol. 37, no. 3, Sept. 2012.

  - Improving column-wise locality
  - Getting the optimal strategy == TSP

- Table Compressors Based on Exploiting Column Dependencies

  Yihan Gao and Aditya Parameswaran, "Squish: Near-optimal compression for archival of relational datasets," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016, KDD '16, p. 1575–1584, Association for Computing Machinery.

  - Column-wise dependencies not utilized

- Entry-wise Predictors

  Pictures: PNG
  HPC Data: fpzip, spdp

  - Applicable only to specific types of 2D data, relies heavily on the pre-assumptions of the datasets

SortComp: A Toy Example

1.50 bit

Job Level	Base	Bonus
Junior	11000	2200
Junior	12000	2400
Junior	10000	2000
Junior	13000	2600
Senior	24000	4800
Junior	14000	2800
Senior	25000	5000

Sorted Table

Job Level	Base	Bonus
Junior	10000	2000
Junior	11000	2200
Junior	12000	2400
Junior	13000	2600
Junior	14000	2800
Senior	24000	4800
Senior	25000	5000

Intra-Column Correlations

Junior → 0
Senior → 1
Column Difference

Job Level	Base	Bonus
0	10000	2000
0	1000	200
0	1000	200
0	1000	200
0	1000	200
1	10000	2000
0	1000	200

Permutation Table

Job Level	Base	Bonus
1	2	2
2	3	3
3	1	1
4	4	4
6	6	6
5	5	5
7	7	7

Inter-Column Correlations

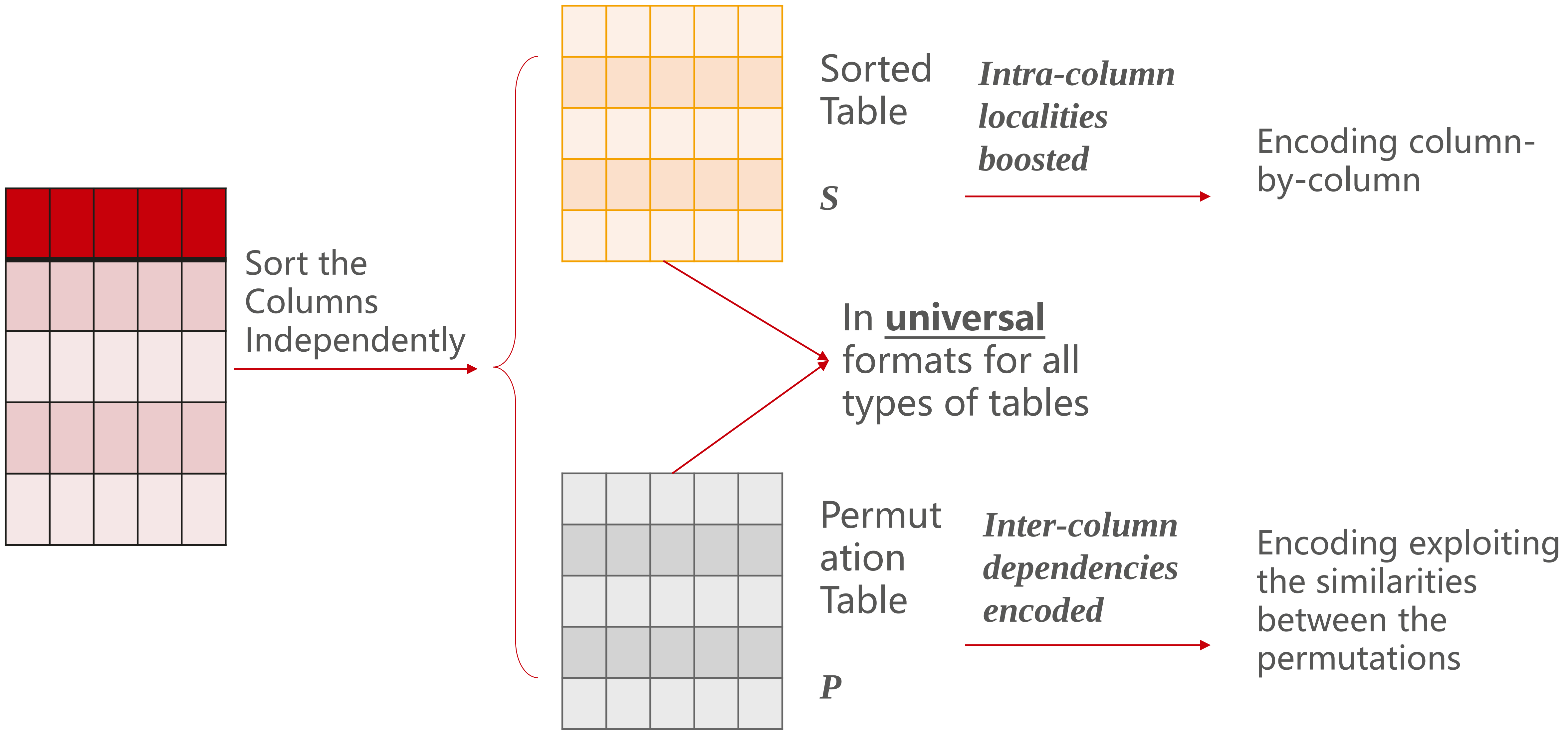
Job Level - Identity
Base - Job Level
Bonus - Base

Job Level	Base	Bonus
0	1	0
0	1	0
0	-2	0
0	0	0
1	0	0
-1	0	0
0	0	0

1.13 bit

HUAWEI

SortComp: Towards A Universal Scheme



HUAWEI

Table w/ One Column Emitted by an i.i.d Source

- SortComp compresses the table to the theoretical limit (ensured by Shannon Entropy) asymptotically.

Lemma 3. *Given an integer sequence* $a = \{a_1, a_2, ...a_n\}$ *with* $1 \leq a_i \leq N, \forall 1 \leq i \leq n$, $N = o(n)$, *drawn from an i.i.d source* π. *Assume permutation vector* $\sigma = \{l_1, l_2, ..., l_n\}$ *sorts* a *to* $b = \{b_1, b_2, ...b_n\}$. *Then as* n *goes to infinity,* b *and* σ *can be encoded with a total length of* $nH(\pi)$ *bits.*

Lemma 2. *Given an integer sequence* $a = \{a_1, a_2, ...a_n\}$ *with* $1 \leq a_i \leq N, \forall 1 \leq i \leq n$, *drawn from an i.i.d stationary source. Denote the average of the drawn sequence as* λ, *i.e.* $\frac{\sum_{i=1}^{n} a_i}{n} = \lambda$, *then it stands:*

Geometric Dist.

$$H(a) \leq 1 + \log \lambda, \quad (1)$$

HUAWEI

Table w/ One Column Emitted by an i.i.d Source

- SortComp compresses the table to the theoretical limit (ensured by Shannon Entropy) asymptotically.

Lemma 3. *Given an integer sequence* $a = \{a_1, a_2, ...a_n\}$ *with* $1 \leq a_i \leq N, \forall 1 \leq i \leq n$, $N = o(n)$, *drawn from an i.i.d source* π. *Assume permutation vector* $\sigma = \{l_1, l_2, ..., l_n\}$ *sorts* a *to* $b = \{b_1, b_2, ...b_n\}$. *Then as* n *goes to infinity,* b *and* σ *can be encoded with a total length of* $nH(\pi)$ *bits.*

↑

Lemma 2. *Given an integer sequence* $a = \{a_1, a_2, ...a_n\}$ *with* $1 \leq a_i \leq N, \forall 1 \leq i \leq n$, *drawn from an i.i.d stationary source. Denote the average of the drawn sequence as* λ, *i.e.* $\frac{\sum_{i=1}^{n} a_i}{n} = \lambda$, *then it stands:*

Geometric Dist.

$$H(a) \leq 1 + \log \lambda, \quad (1)$$

Two Columns

- Compress column-by-column + utilizing inter-column dependencies boost the compression ratio

noise, measurement error, ...

Theorem 1. *Given two columns $A^{(i)}$ and $A^{(j)}$ of table A, assume $A^{(j)} = g(A^{(i)}) + \beta$, where β is a random variable. Then given $P^{(i)}$, $P^{(j)}$ can be encoded with a total length of $n + n\log O(\frac{d(g(A^{(i)}),A^{(j)})}{n})$, where d refers to the Kendall Tau distance.*

- Recall: The information of any permutation P with length n is $nO(\log n)$
- Space save: $O(\log \frac{n^2}{d(g(A^{(i)}),A^{(j)})})$ bits for each entry

Can be small if β is small

Column 1	Column 2
1	15
2	20
3	25
4	27
5	28
6	27
7	30
8	34

Almost in the same order

Contribute to Kendall-Tau distance

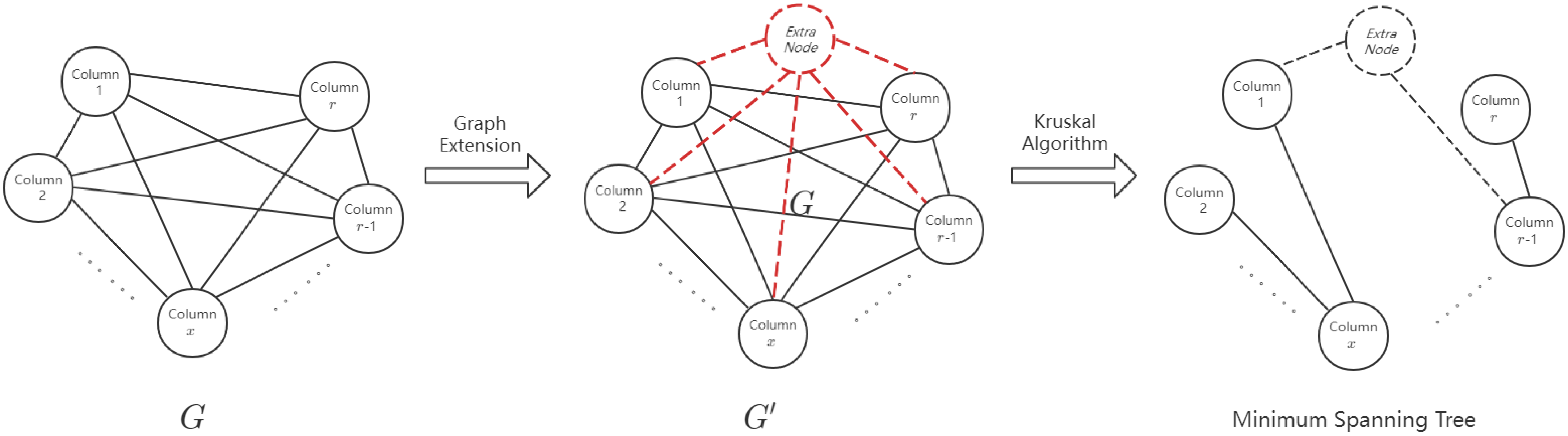
Encoding Routine: (1). Compute the composite permutation $P = P^{(i)^{-1}} \circ P^{(j)}$; (2). Compute the Lehmer code $L = L(P)$; (3). Entropy-encode the Lehmer code L.

HUAWEI

Multi-Column Optimization: A Graph Approach

Table T ⟷ Graph G Column C ⟷ Node v Edge $v_1 v_2$ ⟷ code length of v_2 based on v_1

Extended Graph $G' = G \cup extra\ v'$ Edge $v'v$ ⟷ code length of v



Experiments

Table 1: Experimental Compression Ratios of Different Algorithms

File Name	Gzip Row/Col	zstd Row/Col	*SortComp*	Spartan
Citibike	5.05/5.54	11.62/8.56	**15.61**	NA
Sanfrancisco Salaries	2.92/3.61	4.17/**4.92**	4.75	NA
Sales Data	3.76/7.63	5.67/8.92	**12.87**	NA
911 Calls	4.24/5.45	11.51/9.13	**12.47**	NA
Levels Fyi Salary	4.42/6.21	7.96/9.43	**9.97**	NA
NYC Accidents 2020	4.47/5.99	9.31/9.02	**9.69**	NA
Forest Cover	6.02/9.81	9.15/**18.16**	11.01	10.00
Corel	3.55/3.80	5.07/5.54	**6.41**	3.48

HUAWEI