인-메모리 row store 데이터베이스에서의 비동기적 데이터 압축

Asynchronous row store compression in In-memory Database

Petabyte-scale In-memory Database Lab
정혁진

INTRODUCTION

Data Compression?
Encoding information using fewer bits than original representation.

Why so necessary?
In-memory database loads whole data in memory. Because of memory pressure, fast and compact compression is necessary.

Why row-store?
Column store database already supports compression. **But,** row store has its own advantage, OLTP performance.

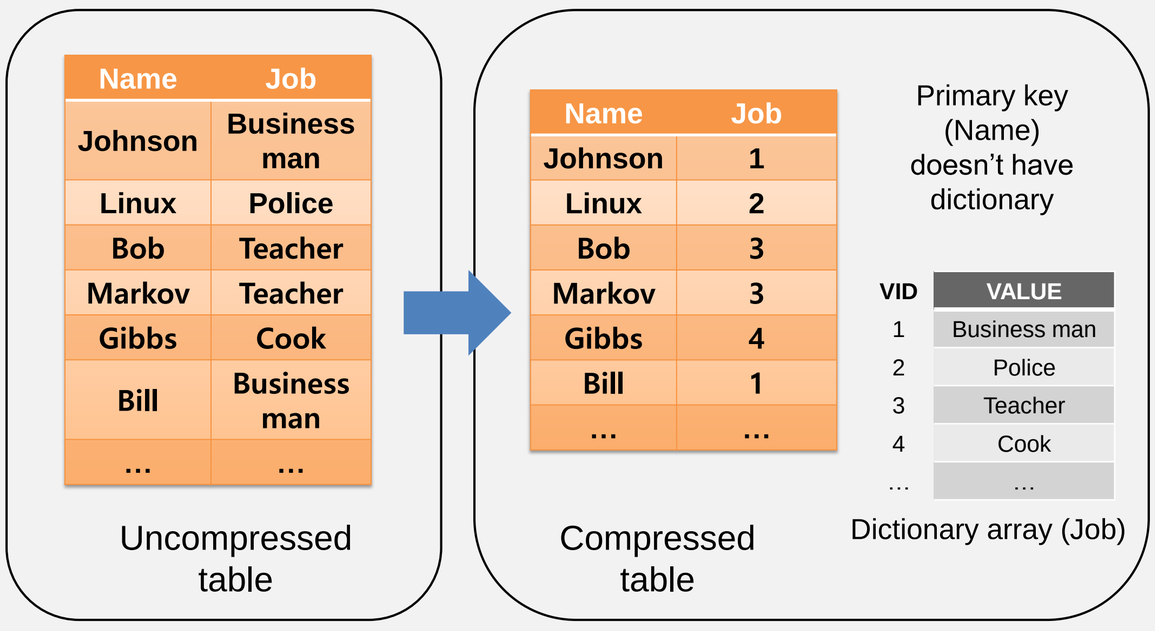
Difficulty
Performance degradation, unified index, evaluation of compression rate, row storage data structure.

Dictionary Encoding

Motivation
Remove data redundancy

Drawback
Dictionary build, update, look-up, Data encoding, decoding

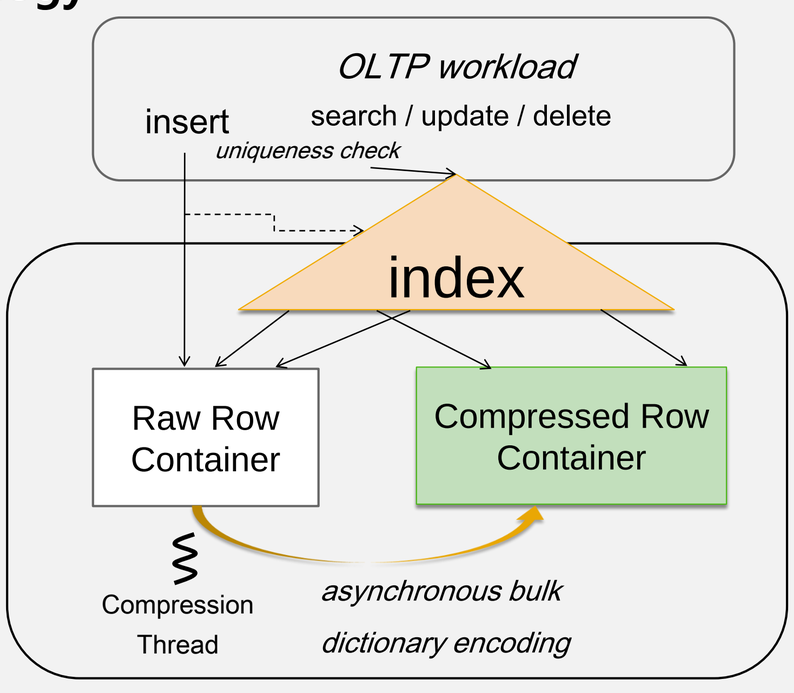
Methodology



Asynchronous Compression

Motivation
Overlap compression time and OLTP time

Methodology



Compression Rate evaluation

Cardinality
- Table cardinality : number of tuples in a relation
- Column cardinality : number of distinct values in a column

Entropy
- Measure for information density
- Entropy = column cardinality / table cardinality

$$\textbf{Compression rate} = \frac{\text{Dictionary size} + \text{Compressed data}}{\text{Original data}}$$

(In case of fixed field size)

$$\textbf{Compression rate} = \text{Entropy} * \text{field size} + \frac{\lceil \log_2 column\ cardinality \rceil}{field\ size}$$

↑ Dictionary ↑ Compressed data

EXPERIMENT

Experiment Environment
CPU : Intel Xeon CPU (40 cores), Memory : 1TB

Memory consumption
TPC-H Scale factor 10(10GB)
(Transaction Processing Performance Council)
Target table : Fact table (LINEITEM. 80% of whole data)

OLTP
SD benchmark (Sales and Distribution benchmark)
A sell-from-stock scenario, which includes the OLTP workloads.
15 Dialog Steps(DS) are repeated.

OLAP
TPC-H SF10 queries which contain compressed table

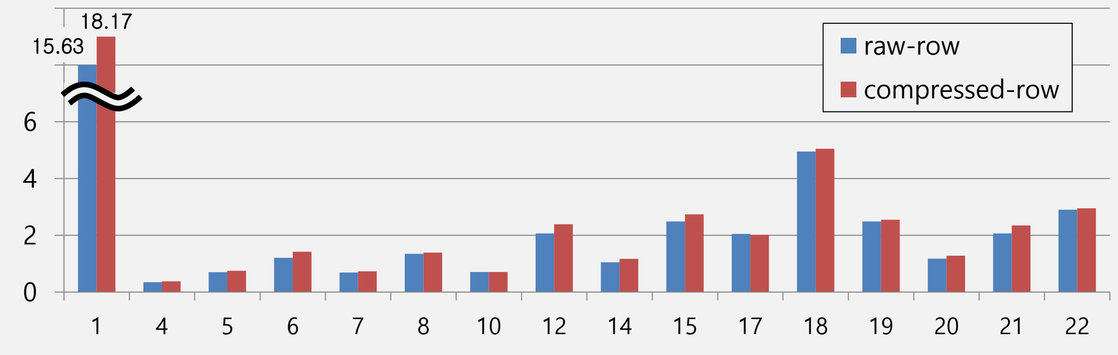
RESULT AND BOTTLENECK

Result
Memory consumption : Compression rate 0.46

OLTP : 9% insertion performance degradation

SD benchmark	Row	Compressed Row	Column
Throughput (DS/min)	220,236	200,214	71,337
Throughput Decrease	0%	-9%	-68%

OLAP : About 8% slower execution of analytic query



Performance Bottleneck
Decoding cost → Reduce decoding count [To-Do]