Big Data Compression Tool using Attribute-based Signatures

Constantinos Costa
Rinnoco Ltd
Limassol, Cyprus
costa.c@rinnoco.com

Panos K. Chrysanthis
Rinnoco Ltd, Limassol, Cyprus
University of Pittsburgh, PA, USA
panos@{rinnoco.com,cs.pitt.edu}

Herodotos Herodotou
Cyprus University of Technology
Limassol, Cyprus
herodotos.herodotou@cut.ac.cy

Marios Costa
Rinnoco Ltd
Limassol, Cyprus
marios.c@rinnoco.com

Efstathios Stavrakis
Algolysis Ltd
Limassol, Cyprus
stathis@algolysis.com

Nicolas Nicolaou
Algolysis Ltd
Limassol, Cyprus
nicolas@algolysis.com

ABSTRACT

This paper introduces COMPASS, a multiple compression tool utilizing attribute-based signatures. COMPASS exploits K-means clustering to select the best compression scheme for different data subsets in a database. The experimental results show that COMPASS significantly reduces disk space usage compared to monolithic methods.

KEYWORDS

big data, signature based, compression, column stores, hybrid store

1 THE COMPASS TOOL

The hypothesis of COMPASS is that multi-scheme data compression is more effective for complex big data by enabling incremental compression and partial decompression. Multi-scheme data compression employs different compression schemes that are more effective for various subsets of data based on their characteristics.

COMPASS breaks down relational data into rows or columns and applies the most suitable compression scheme to individual columns or groups of columns. COMPASS uses K-means clustering to group similar columns together, based on their data values or entropy, and then applies the best compression technique for each cluster. COMPASS selects a compression scheme for an individual column or group of columns by utilizing the database catalog and historical workload information (attribute-based signatures). This work builds on our previous project, SIBACO, which allowed us to observe the basic principles and formulate the technology concept [1].

2 EXPERIMENTAL EVALUATION

This section provides details regarding datasets and techniques used for the preliminary evaluation. In our experimentation, we chose the same readily available compression algorithms (i.e., LZMA, BZIP2, and DEFLATE) from the *zipfile* library, used in SIBACO.

We utilized K-means to group the columns, after scaling the input data using the StandardScaler, and encoding all string columns using the OrdinalEncoder from the *sklearn 1.4.2* library. To validate our proposed ideas and evaluate COMPASS, we conduct the following experiment over two Ubuntu 22.04 server, each featuring 24GB of RAM with Intel(R) Xeon(R) E5-2630 CPU.

Compared Techniques: Our aim in this experimental series is to compare the following four techniques:

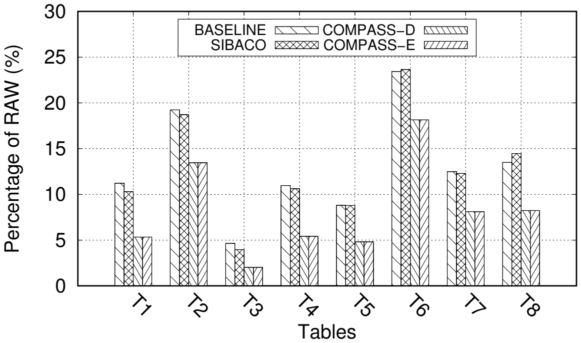

Figure 1: Disk Space for the eight largest tables in the Envmon Database

BASELINE: This baseline technique compresses the data without considering the data characteristics, using the best compression scheme from the *zipfile* library for a given table.

SIBACO: This is our previous technique that employs multiple compression schemes [1].

COMPASS-D: This is our first proposed technique, which uses K-means clustering to group similar columns together to achieve the best compression ratio using multiple schemes.

COMPASS-E: This is our second proposed technique that applies K-means clustering based on the entropy of the columns, which has significantly lower computation complexity than COMPASS-D.

Envmon Dataset: This is a real dataset from the Environmental Monitoring database, developed through the STEAM project (Sea Traffic Management in the Eastern Mediterranean). The database contains primarily environmental, meteorological and oceanographic data. The dataset was collected over the course of three year and has a total size of ~1GB.

Preliminary Results (Fig. 1): Across all tables, COMPASS-D and COMPASS-E consistently demonstrate the most efficient disk space reduction, resulting to 2–18.2% of the original RAW size. In contrast, BASELINE and SIBACO methods reduce the disk space requirements, 4.6–23.4% and 4–23.6% of the original RAW size, respectively. COMPASS-D and COMPASS-E outperform BASELINE and SIBACO by more than 22% in the worst case and 56% in the best case.

REFERENCES

[1] Constantinos Costa, Panos K. Chrysanthis, Marios Costa, Efstathios Stavrakis, and Nicolas Nicolaou. 2023. Towards a Signature Based Compression Technique for Big Data Storage. In *39th IEEE International Conference on Data Engineering, ICDE - Workshops*. 100–104. https://doi.org/10.1109/ICDEW58674.2023.00022

The project is implemented under the programme of social cohesion "THALIA 2021-2027" co-funded by the European Union, through Research and Innovation Foundation.