Analyzing Data Properties using Statistical Sampling Techniques

Illustrated on Scientific File Formats and Compression Features

Julian Kunkel

Deutsches Klimarechenzentrum (DKRZ)

DKRZ DEUTSCHES KLIMARECHENZENTRUM

Abstract

Understanding the characteristics of data stored in data centers helps computer scientists identifying the most suitable storage infrastructure to deal with these workloads. For example, knowing the relevance of file formats allows optimizing the relevant file formats but also helps in a procurement to define useful benchmarks.

Existing studies that investigate performance improvements and techniques for data reduction such as deduplication and compression operate on a small set of data. Some of those studies claim the selected data is representative and scale their result to the scale of the data center. One hurdle of evaluate novel schemes on the complete data is the vast amount of data stored and, thus, the resources required to analyze the complete data set. Even if this would be feasible, the costs for running many of those experiments must be justified.

This poster investigates stochastic sampling methods to compute and analyze quantities of interest on file numbers but also on the occupied storage space. It is demonstrated that scanning 1% of files and data volume is sufficient on DKRZ's supercomputer to obtain accurate results. This not only speeds up the analysis process but reduces costs of such studies significantly.

Contributions of this poster are:

1. investigation of the inherent error when operating only on a subset of data
2. presentation of methods that help future studies to mitigate this error
3. illustration of the approach with a study for scientific file types and compression

The Problem

When analyzing properties of data (quantities of interest), they vary across the file system. Conducting reliable studies on data requires to scan large quantities of data. Scanning a subset of data without proper sampling techniques, can lead to wrong conclusions.

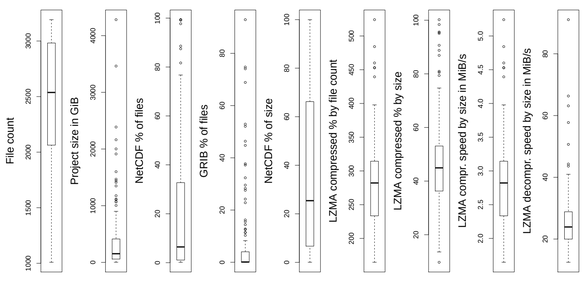


Fig. 1: Variability across 125 projects – each point represents the mean of one project.

Example Study

For demonstrating the approach, several quantities of interest are investigated:

- Distribution of file sizes
- Used scientific file formats (proportions)
- Compression ratio ZIP, GZIP, BZIP2, LZMA
- (De)-compression speed

Arithmetic means are computed based on file count and occupied file size.

Analyzed Data

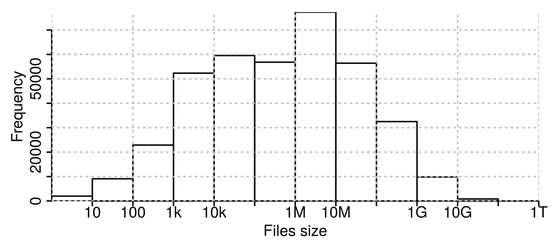
- Lustre file system of Mistral
- 380k files out of 320 million (0.12%)
- 53.07 TiB of data out of 12 PiB (0.44%)

Approach for Scanning

- Concurrently scan accessible files of each project directory using find in individual lists.
- Create a random sample of 10k files of each project; merge them into a single file list.
- Create a permutation of the file list; partition the result into one list for each thread.
- Distribute/run threads on different nodes, each processes one list sequentially.
- After two weeks, threads are terminated; resulting data is ingested into a SQLlite DB.

Exploring Analyzed Data

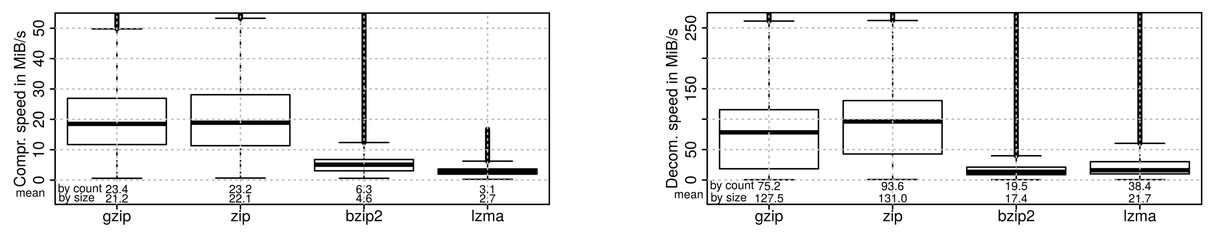
Distribution of file sizes



(a) Histogram

(b) Cumulative file sizes (y-axis in log scale)

Compression/decompression speed per file

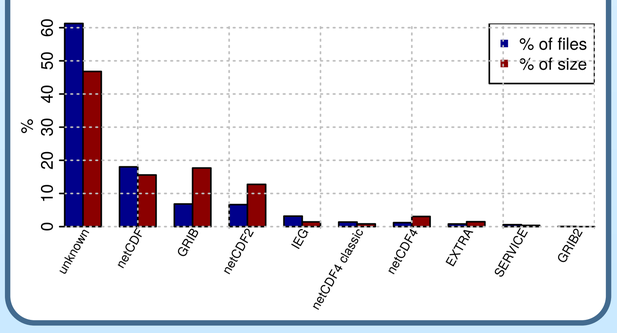


(c) Compression

(d) Decompression

The arithmetic means (also computed on file size) are shown as text under the plot.

Scientific File Formats



Analysis: Conclusions

- Investigated quantities may help to decide:
  - For which data size to optimize
  - Which compression scheme to choose
  - For which file formats start optimizing
- Projects exhibit a high variability
- Computing arithmetic mean over file count and by occupied file size differs; the reason is the heavy-tailed distribution of file sizes
- We need reliable sampling methods not only for the typical file but also across file size

Stochastic Sampling of Data

The strategy for selecting files appropriately and compute proportions or means of variables by file count, e.g., arithmetic mean, is simple, while the strategy to weight means by file size is non-trivial.

Strategy to compute by file count

1. enumerate all files on the storage system
2. create a simple random sample, i.e., choose a number of files
3. determine the quantities of interest

Strategy to compute by file size

1. enumerate all files; determine their sizes
2. pick a random sample (with replacement) based on the probability defined by filesize/totalsize
3. determine quantities of interest for all unique files
4. compute the mean among all samples (no additional weighting of file size)

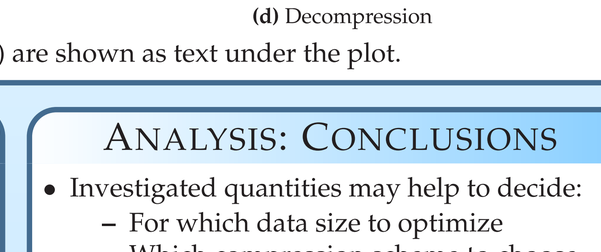
Convergence

To understand convergence, the sampling methods are applied on the population of scanned files, simulation is used – it is applied 100 times where the estimated mean is determined. The boxplots in Figure 2 and 3 show the results of several quantities of interest for analyzing 0.1%, 1% and 5% of files and for 256 and 4096 samples when weighting occupied size. It can be seen that the variance of applying the strategy converges to the true mean and by scanning 1% of files and drawing 4096 samples for computing by file sizes yields good results.

Fig. 2: Sampling by file count

(a) Sampling 316 = 0.1% of files)

(b) Sampling 3164 = 1% of files)



(c) Sampling 6329 = 5% of files)

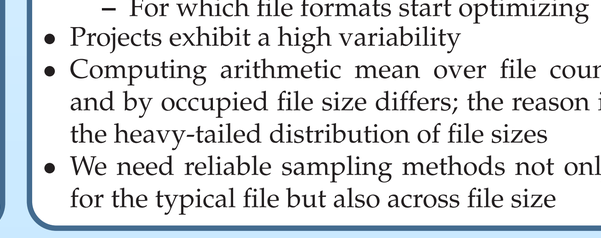
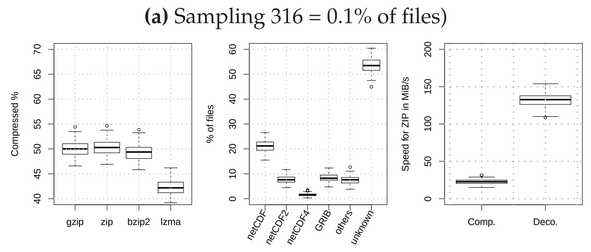
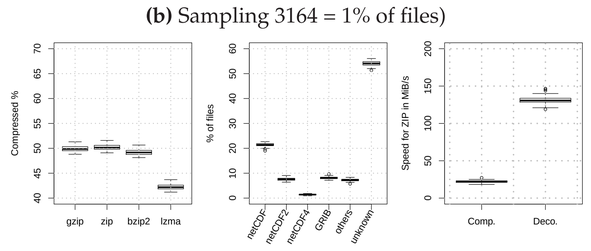


Fig. 3: Correct sampling; weights according to file size

(a) Drawing 256 samples with replacement



(b) Drawing 4096 samples with replacement



Understanding the Error and Necessary Sample Size

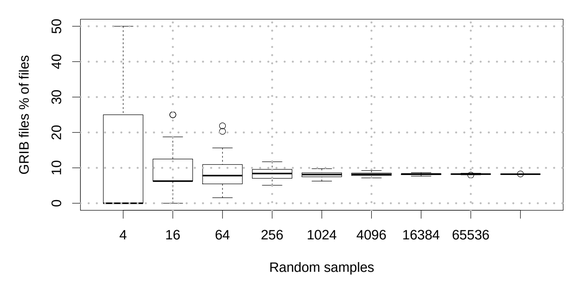
To select the number of observations that allows inference about the population, statistics knows methods for determining sample size. Before determining the sample size, the acceptable error based on the used analysis method is chosen and a confidence level such as 95% is decided.

For estimating proportions there are approaches that work regardless of probability distribution: With Cochran's sample size formula, to achieve an error bound of $\pm 5\%$ and $\pm 1\%$, roughly 400 and 10000 samples are needed, respectively [1]. Note that the sample number does not increase even for large population sizes. Estimating a continuous variable such as the arithmetic mean compression ratio or performance is complex, as sampling ratio is based on the expected distribution of values that is not normally distributed. All methods have in common that they determine the mean value not by weighting occupied storage space.

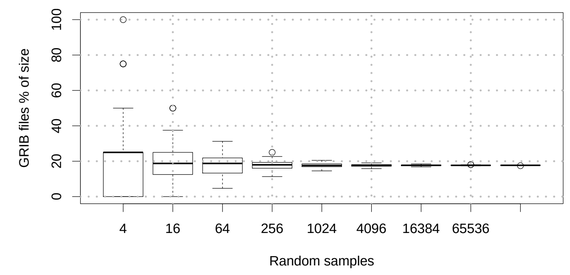
Understanding Robustness

To understand the convergence better, for an increasing number of samples a simulation has been conducted for the proportion of GRIB files. The last experiment illustrates the problem when trying to analyze means proportional to file size but sampling by file count. This simple approach shows a suboptimal convergence behavior and is not reliable.

Compute by file count

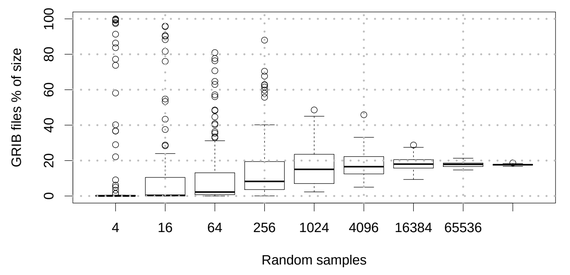


Compute by file size



(c) Sampling (with replacement) with a probability by size

Computation for file size by file count



(d) Sampling by file count (this is not appropriate!)

Summary and Outlook

- The strategies described allow to capture representative samples
- This allows to reliably conduct studies of data characteristics
- Scanning about 1% of available files and occupied capacity yields already good accuracy
- Applying a wrong sampling strategy leads to unreliable results
- Several interesting characteristics of the data could be deducted
- We will work on tools to automatize this process and automatically quantify the error

Related work

[1] J. Kotrlik and C. Higgins, "Organizational Research: Determining Appropriate Sample Size in Survey Research Appropriate Sample Size in Survey Research," Information technology, learning, and performance journal, vol. 19, no. 1, 2001, p. 43.

[2] S. Lakshminarasimhan, N. Shah, S. Ethier, S.-H. Ku, C.-S. Chang, S. Klasky, R. Latham, R. Ross, and N. F. Samatova, "ISABELA for Effective in Situ Compression of Scientific Data," Concurrency and Computation: Practice and Experience, vol. 25, no. 4, 2013, pp. 524–540.

[3] U. Schulzweida, L. Kornblueh, and R. Quast, "CDO User's guide: Climate Data Operators Version 1.6. 1," 2006.

[4] A. Tursunalieva and P. Silvapulle, "Estimation of Confidence Intervals for the Mean of Heavy Tailed Loss Distributions: A Comparative Study Using a Simulation Method," 2009.

[5] K. Jin and E. L. Miller, "The Effectiveness of Deduplication on Virtual Machine Disk Images," in Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference. ACM, 2009, p. 7.

[6] J. Lofstead, M. Polte, G. Gibson, S. Klasky, K. Schwan, R. Oldfield, M. Wolf, and Q. Liu, "Six Degrees of Scientific Data: Reading Patterns for Extreme Scale Science IO," in Proceedings of the 20th international symposium on High performance distributed computing. ACM, 2011, pp. 49–60.

[7] S. D. Legesse, "Performance Evaluation of File Systems Compression Features," Master's thesis, University of Oslo, 2014.

[8] A. Zuck, S. Toledo, D. Sotnikov, and D. Harnik, "Compression and SSDs: Where and How?" in 2nd Workshop on Interactions of NVM/Flash with Operating Systems and Workloads (INFLOW 14). Broomfield, CO: USENIX Association, Oct. 2014. [Online]. Available: https://www.usenix.org/conference/inflow14/workshop-program/presentation/zuck

[9] N. Hübbe and J. Kunkel, "Reducing the HPC-Datastorage Footprint with MAFISC – Multidimensional Adaptive Filtering Improved Scientific data Compression," Computer Science - Research and Development, 05 2013, pp. 231–239. [Online]. Available: http://link.springer.com/article/10.1007/s00450-012-0222-4

[10] D. Meister, J. Kaiser, A. Brinkmann, M. Kuhn, J. Kunkel, and T. Cortes, "A Study on Data Deduplication in HPC Storage Systems," in Proceedings of the ACM/IEEE Conference on High Performance Computing (SC). IEEE Computer Society, 11 2012.

[11] M. Kuhn, K. Chasapis, M. Dolz, and T. Ludwig, "Compression By Default – Reducing Total Cost of Ownership of Storage Systems," 06 2014. [Online]. Available: http://www.isc-events.com/isc14_ap/presentationdetails.htm?t=presentation&o=253&a=select

[12] J. Kunkel, M. Kuhn, and T. Ludwig, "Exascale Storage Systems – An Analytical Study of Expenses," Supercomputing Frontiers and Innovations, 06 2014, pp. 116–134. [Online]. Available: http://superfri.org/superfri/article/view/20

[13] R. B. Dell, S. Holleran, and R. Ramakrishnan, "Sample size determination," Ilar Journal, vol. 43, no. 4, 2002, pp. 207–213.

Acknowledgements: I thank Charlotte Jentzsch for the fruitful discussions.