Synthetic Data Generation for S3 traffic data using Generative Models

Aaditya Parthasarathy, Aditi Kharkwal, Ameya Bhamare, Neel Shah, Soham Butala

Motivation

As AWS evaluates new hard drives for S3, synthetic data is key for accurately predicting future traffic patterns. Traditional testing methods often struggle to capture the complexity of future hard drive traffic patterns. Synthetic data offers a solution by approximating these patterns, a comprehensive evaluation for new hard drives' performance.

Problem Statement

Synthetic Data Generation based on:

- **Realism**: mirrors the complex patterns of actual S3 traffic
- **Historical Data Constraints**: utilizing parameters like time range, disk volume, container groupings, etc
- **Generative Model Application**: Successfully applying generative models like GANs to produce realistic, synthetic S3 traffic data

Data Profile

- 271 million rows - S3 transactional data
- 13 columns - 11 categorical, 2 numerical
- Encrypted

Data Generation Methods

1. **TabGAN:**
   - An evolution of CTGAN for tabular datasets
   - Handling mixed data types and addressing distribution imbalances
   - Challenges with generating negative values in sparse data, lacks constraints for sample generation

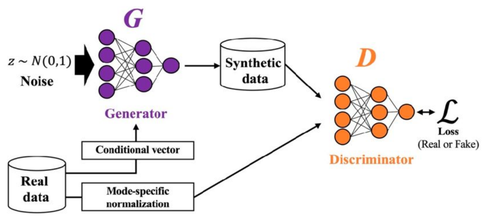
2. **PAR:**
   - Probabilistic AutoRegressive model for learning multivariate time series data
   - Conditionally generate time series based on associated entity's properties
   - Failed since PAR needs features dependant on time, not suitable for S3 data

3. **TimeGAN:**
   - GAN architecture, tackles the distinctive challenges inherent in time series data
   - prioritizes capturing temporal patterns
   - It does not support categorical features, making it unusable for our use case

4. **DoppelGANger:**
   - Directly captures the statistical properties of time series data
   - Struggles with multiple categorical features, fails to capture time trends for non-periodic datasets, affecting its translation to categorical columns

CT-GAN

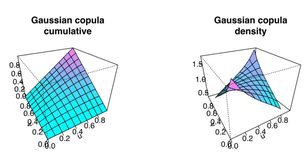
Conditional Tabular Generative Adversarial Network, is a generative model designed to synthesize realistic tabular data by capturing the conditional relationships present in the input dataset.



Gaussian Copula

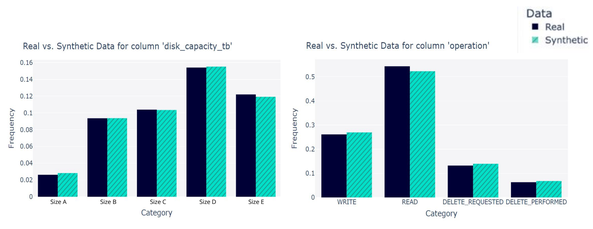
Statistical model that simplifies the analysis of dependence between random variables, transforming them into a joint distribution with uniform margins using gaussian copula function.

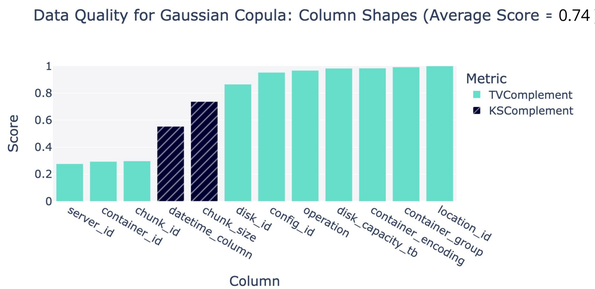
$$C_R^{\text{Gauss}}(u) = \Phi_R\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right)$$



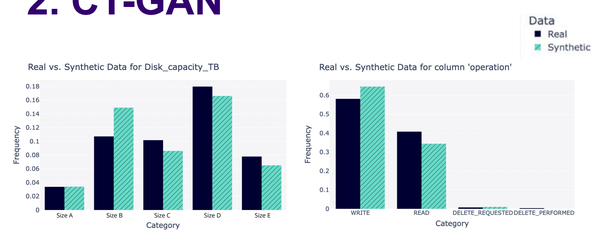
Results

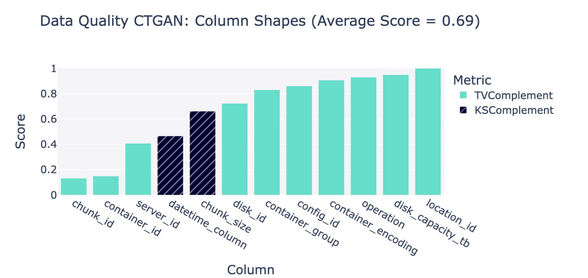
1. Gaussian Copula





2. CT-GAN





3. Gaussian Copula vs CT GAN

Comparing normalized frequency of occurrence of features in training data versus synthetic data across all columns

Column	Gaussian Copula	CT-GAN
Disk Capacity	99.62%	71.81%
Operations	96.93%	84.16%
Container Group	99.87%	77.07%

Conclusions

- Gaussian Copula and CTGAN performed best compared to other models for this use case
- Gaussian copula is 14x faster for 100x more data than CT-GAN for the same compute power - training and prediction
- Gaussian Copula exceeds CTGAN's data quality by 8% for 100X lesser data

References

- Xu, L., Skoularidou, M., Cuesta-Infante, A., & , K. (2019). Modeling Tabular Data using Conditional GAN. In 33rdnference on Neural Information Processing Systems (NeurIPS 2019)
- Ashrapov, Insaf. "Tabular GANs for uneven distribution." arXiv preprint arXiv:2010.00638 (2020).
- Lin, Z., Jain, A., Wang, C., Fanti, G., & Sekar, V. (2019). Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions.