EuroCC@Turkey

https://eurocc.truba.gov.tr/

This document is prepared by EuroCC@Turkey for EuroCC under GA NO 951732

CASE STUDY REPORT
Synthetic IoT Data Generation using HPC Infrastructure

NCC Partner	*METU (Middle East Technical University)*
Company*	*TÜPRAŞ – https://www.tupras.com.tr/* *Mert Onuralp GÖKALP - mertonuralp.gokalp@tupras.com.tr* *Tayfun EYLEN – Tayfun.Eylen@tupras.com.tr*
Expert	*Altan KOÇYİĞİT – kocyigit@metu.edu.tr* *Erhan EREN – ereren@metu.edu.tr* *Kerem NAZLIEL – knazliel@metu.edu.tr* *TRUBA: Kerem KAYABAY – kerem.kayabay@tubitak.gov.tr*
Start & End Date	*14.02.2022 – 16.11.2022*
Approved by	*NCC Project Management Team*

***Company** accepts that the Case Study Report is shared with the EuroCC Project and the community through the EuroCC@Turkey awareness creation activities and platforms.

1. Problem Identification

Turkey Petroleum Refineries Corporation (TÜPRAŞ) operates four refineries in Turkey, controlling most of Turkey's crude oil refining capacity. Many heterogeneous units in the field produce time-series Internet of Things (IoT) data. Among the factors that cause heterogeneity are various brands, tasks, degradation levels, and configurations. These units generate more detailed alarm data (e.g., frequency, action) or process data (e.g., key-value pairs) logged in lower granularity. The company wants to utilize these datasets for multiple purposes, such as alarm filtering and root-cause analysis. However, the industrial and heterogeneous nature of these units, varying granularities in datasets, and missing periods cause problems in data analysis. Within this project's scope, TÜPRAŞ will try to generate synthetic IoT data leveraging HPC resources to overcome issues caused by data quality.

2. First Suggestion

Deep learning-based Generative Adversarial Networks (GANs) are a possible synthetic data generation approach. GANs deliver outstanding results, particularly for image-to-image translation tasks and realistic but fake photogeneration. However, recent studies also explore using GANs for generating synthetic time-series data. Training GANs can be computationally intensive and iterative, making the problem a good fit for an HPC infrastructure. In this case study, TÜPRAŞ will try the HPC infrastructure as a service for the first time. Therefore, to mitigate potential data protection issues, we want to try out the HPC infrastructure with simulated data sets instead of actual data. TÜPRAŞ will also perform model training with real data using local hardware.

The *updated* project plan is as follows:

Timeline	**Tasks and Milestones**
11.03.2022	Orientation on the TRUBA platform Identification of units with high quality data for real data sets
16.06.2022	Identification of suitable technologies and methodologies for pre-processing, training, inference, and evaluation *Milestone*: Simulation data ready for further analysis
30.09.2022	Training the GAN architecture on local hardware using real data set *Milestone*: First results obtained on local hardware with real data set
16.11.2022	Preparation of final report *Milestone*: Final report and future work

3. Solution Stage – I

In the project's first stage, the researchers set up their TRUBA account and became familiar with using Slurm, Python Libraries, and Frameworks on the TRUBA infrastructure. Next, the goal is to simulate the alarm and process data of TÜPRAŞ's production unit. Firstly, a literature review is conducted, and the most suitable production process is determined as the Tennessee Eastman Process. Then, the necessary software packages to simulate this process

are determined. MATLAB and the dedicated Tennessee Eastman Process [1] library are the decided technology to run this simulation. Currently, we are working on simulating process and alarm data for different scenarios. Process variables consist of the operating conditions such as temperature, pressure, the change in raw materials, byproducts, and outputs. Alarm data consists of various alarms arising from simulated materials and operating conditions. After optimizing data simulation in MATLAB, we plan to transfer the generated data to TRUBA resources and apply deep generative models to produce meaningful alarm data.

4. Solution Stage – II

At this stage, we decided to try an open-source Python package [2] that implements several GAN architectures for tabular and sequential data. It implements TimeGAN [3], which can capture the potentially complex dynamics of multivariate sequential variables across time. To test the performance of TimeGAN across a variety of time-series data (i.e., sines, stocks, energy, and events), the authors apply t-SNE and PCA analyses on both the original and synthetic datasets to visualize how closely the distribution of generated samples resembles the original in 2-dimensional space. The authors also train a prediction model by optimizing a 2-layer LSTM to see if the sampled data inherits the predictive characteristics of the original. While the former (i.e., visualization) provides a qualitative assessment, the latter (i.e., predictive performance) provides a quantitative evaluation of the generated samples.

5. Results and Achievements

We started by training TimeGAN on local hardware using actual data set for performance comparison before using the simulated data set on TRUBA infrastructure. Figure 1 shows the visualizations with PCA and t-SNE, which demonstrate that the distribution of generated samples covers the actual data in 2-dimensional space, with the former performing better. Furthermore, we fit two linear regression models using real and synthetic data to perform a quantitative assessment, which we use to obtain predictions on the original data set. The MAE score for the real dataset predictions achieves 0.08, which is 0.04 for the synthetic data. The quantitative assessment shows that the synthetic dataset can be valuable for predictive purposes.

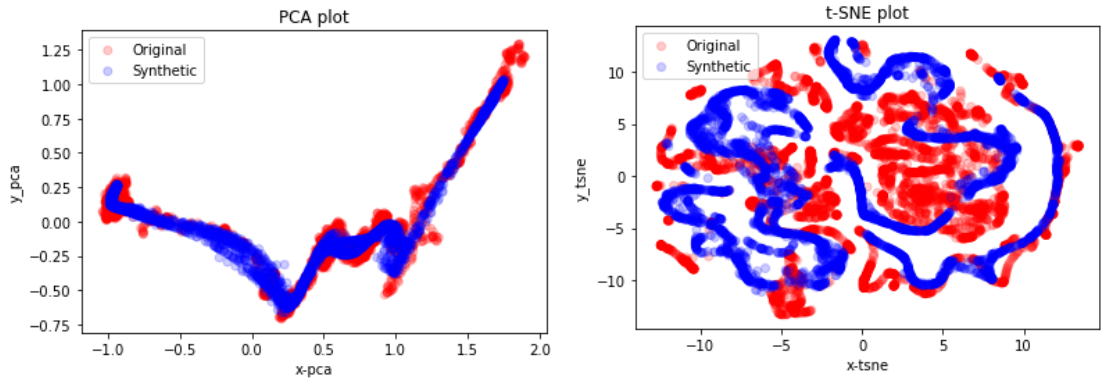

Figure 1: PCA and t-SNE visualizations for real and synthetic data

After testing TimeGAN on local hardware, we transferred the code to the TRUBA infrastructure to use the simulated data set on the HPC infrastructure. Even though we could train GAN network on the HPC infrastructure, we couldn't complete the evaluations we intended in the scope of this case study. We will continue this investigation academically to benchmark local performance with GPU acceleration (e.g., P100 and V100). The GPUs on the TRUBA infrastructure will enable us to increase the size of the training dataset, which can also improve performance.

This case study enabled TÜPRAŞ to focus on data-centric operations rather than model-centric operations by generating synthetic multivariate time-series data using generative models. The generated datasets can significantly improve the data quality, which the data scientists can utilize for multiple tasks, including alarm filtering and root cause analysis. Before this case study, TÜPRAŞ did not explore computationally-intensive tracks, unaware of the available HPC services. We believe the case study carried TÜPRAŞ to the HPC-ready level, presenting an opportunity to study the data-centric AI trend, providing awareness and orientation for HPC services, and initiating collaborations toward successful technology adoption.

References

[1] A. Bathelt, N. L. Ricker, and M. Jelali, “Revision of the Tennessee Eastman Process Model,” *IFAC-PapersOnLine*, vol. 48, no. 8, pp. 309–314, Jan. 2015, doi: 10.1016/j.ifacol.2015.08.199.

[2] “YData Synthetic.” YData, Aug. 25, 2022. Accessed: Aug. 26, 2022. [Online]. Available: https://github.com/ydataai/ydata-synthetic

[3] J. Yoon, D. Jarrett, and M. van der Schaar, “Time-series Generative Adversarial Networks,” in *Advances in Neural Information Processing Systems*, 2019, vol. 32. Accessed: Nov. 14, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/c9efe5f26cd17ba6216bbe2a7d26d490-Abstract.html