4.3 Data Synthesis Overhead

A viable approach to creating synthetic data for accurate performance modeling must not impose excessive overhead for data creation. We measured the overhead of creating synthetic training data on the hardware described in Section 4.1.

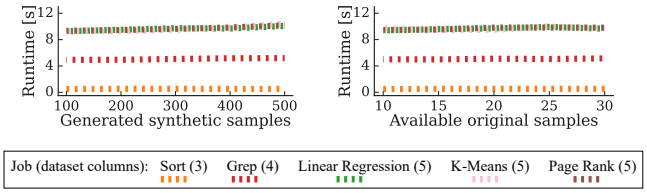

Figure 4: Overhead for creating synthetic data for different Spark job performance datasets.

In Figure 4, we see that for performance datasets containing the runtime and runtime-influencing factors of typical Spark jobs, this overhead was measured to be approximately between half a second and ten seconds. We observe that the computational cost of synthesizing data does not significantly increase with an increase in the amount of sampled synthetic data or the number of available samples in the original dataset. Rather, the findings suggest that the primary computational effort arises from processing each attribute, i.e., column, in the original dataset. In the case of DataSynthesizer, this part is conducted by the DataDescriber component.

4.4 Discussion

We will now discuss the experimental evaluation's results in terms of the practical implications for our approach's viability.

First, we observed that it is feasible to generate substantial quantities of synthetic data without compromising the model's accuracy. This implies that we can achieve privacy not just by modifying the content of each data point, but also by creating arbitrarily large amounts of synthetic data, thereby concealing the actual quantity of processed jobs.

Then, it has been observed that the model accuracy gap when using synthetic data is lowest when the quantity of original data points is low. In instances where publicly shared training data points are unavailable or rare, the introduction of synthetic data can have a significant positive effect on the model accuracy of collaborators. Consequently, sharing synthetic data is particularly advantageous in the early stages of a training data sharing initiative.

Finally, the computational overhead of generating synthetic performance data has been shown to range in seconds for performance datasets of typical Spark jobs on typical consumer hardware. This low amount of time should not discourage collaborators from generating and sharing synthetic data.

5 CONCLUSION

In summary, this paper has explored how differential privacy via data synthesis can facilitate the sharing of runtime data for performance modeling of data analytics workloads in a privacy-preserving manner. Our initial method has demonstrated an acceptable trade-off between model prediction accuracy and data privacy. Especially in cases where there is limited available performance data overall, the accuracy of collaborators' performance models can be significantly improved through the use of shared synthetic training data samples. Further, the data synthesis has been shown to induce low computational overhead.

In the future, we will investigate alternative approaches to ensure privacy when sharing performance metrics of data analytics workloads. Moreover, we hope our short paper also inspires further research by others in the same direction.

ACKNOWLEDGMENTS

This work has been supported through a grant by the German Research Foundation (DFG) as "C5" (grant 506529034).

REFERENCES

[1] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. 2014. Machine Learning Classification over Encrypted Data. *Cryptology ePrint Archive* (2014).
[2] Paris Carbone, Asterios Katsifodimos, Stephan Ewen, Volker Markl, Seif Haridi, and Kostas Tzoumas. 2015. Apache Flink: Stream and Batch Processing in a Single Engine. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 36, 4 (2015).
[3] Haokun Fang and Quan Qian. 2021. Privacy Preserving Machine Learning with Homomorphic Encryption and Federated Learning. *Future Internet* 13, 4 (2021).
[4] Oded Goldreich. 1998. Secure Multi-Party Computation. *Manuscript. Preliminary version* 78, 110 (1998).
[5] Chin-Jung Hsu, Vivek Nair, Vincent W Freeh, and Tim Menzies. 2018. Arrow: Low-level Augmented Bayesian Optimization for Finding the Best Cloud VM. In *ICDCS '18*. IEEE.
[6] Muhammed Tawfiqul Islam, Shanika Karunasekera, and Rajkumar Buyya. 2021. Performance and Cost-Efficient Spark Job Scheduling Based on Deep Reinforcement Learning in Cloud Computing Environments. *IEEE Transactions on Parallel and Distributed Systems* 33, 7 (2021).
[7] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine* 37, 3 (2020).
[8] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Computing Surveys* 54, 2 (2021).
[9] Haoyue Ping, Julia Stoyanovich, and Bill Howe. 2017. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In *SSDBM '17*. ACM.
[10] Debbie Rankin, Michaela Black, Raymond Bond, Jonathan Wallace, Maurice Mulvenna, Gorka Epelde, et al. 2020. Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. *JMIR Medical Informatics* 8, 7 (2020).
[11] Dominik Scheinert, Alireza Alamgiralem, Jonathan Bader, Jonathan Will, Thorsten Wittkopp, and Lauritz Thamsen. 2021. On the Potential of Execution Traces for Batch Processing Workload Optimization in Public Clouds. In *Big Data '21*. IEEE.
[12] Dominik Scheinert, Philipp Wiesner, Thorsten Wittkopp, Lauritz Thamsen, Jonathan Will, and Odej Kao. 2023. Karasu: A Collaborative Approach to Efficient Cluster Configuration for Big Data Analytics. In *IPCCC '23*. IEEE.
[13] Navoda Senavirathne and Vicenç Torra. 2020. On the Role of Data Anonymization in Machine Learning Privacy. In *TrustCom '20*. IEEE.
[14] Shivaram Venkataraman, Zongheng Yang, Michael Franklin, Benjamin Recht, and Ion Stoica. 2016. Ernest: Efficient Performance Prediction for Large-scale Advanced Analytics. In *NSDI '16*. USENIX.
[15] Jonathan Will, Lauritz Thamsen, Dominik Scheinert, Jonathan Bader, and Odej Kao. 2021. C3O: Collaborative Cluster Configuration Optimization for Distributed Data Processing in Public Clouds. In *IC2E '21*. IEEE.
[16] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, Ion Stoica, et al. 2010. Spark: Cluster Computing with Working Sets. *HotCloud* 10, 10 (2010).