Devron

Devron: A System for Privacy-Preserving Federated Learning

Three market forces are heavily influencing the future of our digital ecosystem: (1) data is increasingly voluminous, fragmented, and siloed across various administrative domains, (2) data processing and machine learning are transforming every aspect of society, and (3) there is an increasing number of data privacy, data residency, data sovereignty, and geo-proximity regulations. Therefore, enabling data science across multiple data sources without moving data and which respects the privacy requirements of each data source is needed to unlock the value potential of data and enable organizations to realize the potential machine learning, AI, and data science efforts. Devron is that platform. Its specifics are described in this document.

System Description

Devron enables privacy-preserving machine learning and, more broadly, data science, over distributed— horizontally and/or vertically split—datasets (i.e., similar or different data). Devron includes (1) an infrastructure amenable to rigorous security and privacy quantification, (2) a hybrid combination of privacy-enhancing technologies (from partially homomorphic encryption, synthetic data generation, differential privacy, and k-anonymity style measures) to ensure strong privacy of data, (3) proprietary algorithms to enable machine learning over homogeneous and heterogeneous data sources, and (4) an analysis of the privacy utility trade-offs involved in the system.

Devron is built on bringing code to data and maintains a strict separation between the two. To implement our product, we build upon federated learning, an emerging privacy-enhancing technology. Data never moves, remains under the sole control of the data owner, and can only be analyzed through a set of valid "jobs." Devron enables every step of the data science pipeline in a privacy-preserving manner – exploratory data analysis (EDA), iterative model training, and model serving.

Unique Value Proposition

Among Devron's valuable offerings is the ability to extract value from multiple data sources without centralizing the data and without exposing the raw values. Consider a simple example of two data sources—(e.g., a payment gateway with information such as which account transacts with which account) and another holding bank data (e.g., account information, duration, etc.). Currently, the payment gateway is limited to using its data in detecting fraudulent transactions (say X% recall on fraudulent transactions). With Devron, the bank dataset can be used to improve the performance of this model, i.e., to achieve a recall of Y% where $Y \geq X$. The approach, techniques, and mathematics are not specific to any dataset or scenario and generally apply to any multi-dataset scenario, be they residing in different systems, on different clouds, across different jurisdictions, or owned by different organizations.

We achieve this by using privacy-preserving algorithms and cryptographic techniques combined with state-of-the-art security and privacy practices. In addition, we use a combination of an alignment routine (that privately maps relevant PIIs between datasets) and an assortment of local and global machine learning training to achieve boosted performance. Our research shows that each approach (for different datasets and problems) demonstrates the ability to learn from multiple data sources while preserving privacy. We thus crucially unlock improved performance gains than training on a single data source.

Summary of Privacy Posture

Devron uses partially homomorphic encryption to align PII across data sources, synthetic data to allow the EDA without exposing the raw data, differential privacy to hide metadata information and statistics critical for effective data analytics, and supports additional k-anonymity style privacy measures to prevent privacy leakages from computation on small amounts of data. We use a defense-in-depth approach—a combination of cloud security, best industry practices, and state-of-the-art cryptographic protocols and PETs to enable a data science experience that simultaneously offers consistent data utility while upholding strong security and privacy guarantees.

The combination of these capabilities allows users to unlock previously-inaccessible datasets, shorten the time to access data, accelerate experimentation and time to insight, reduce the risk and overhead of moving data, and produce analytics of greater accuracy and effectiveness.