Benefits and Challenges of Decentralization in Data Systems: Opportunities for Data Management Research

Ruben Mayer
University of Bayreuth
Bayreuth, Germany
ruben.mayer@uni-bayreuth.de

ABSTRACT

Decentralization promises to overcome many of the shortcomings of centralized data systems: It puts clients in control of their own data and the processing that they perform on it, without requiring them to trust a central entity. However, decentralization faces many challenges. While the problems of efficiency and security have been in the focus of the data management community, in this keynote talk, I emphasize the need for research on other challenges of decentralized data systems, namely, governance and responsibility. Based on examples from blockchain and federated learning systems, I argue that these challenges require more attention from the data management community and open many new research opportunities.

VLDB Workshop Reference Format:
Ruben Mayer. Benefits and Challenges of Decentralization in Data Systems: Opportunities for Data Management Research. VLDB 2024 Workshop: Foundations and Applications of Blockchain.

1 MOTIVATION

The classical design of data systems is *centralized*: All processing units and all data are controlled by a single entity. This way, the entire data pipeline can be optimized by that entity, allowing for the design of highly efficient systems. In particular, there is no need to coordinate the processing across multiple stakeholders; instead, all necessary decisions are made by the central entity that is in control. This does not mean that the processing units themselves cannot form a distributed system. The difference between decentralization and distribution is based on the question who controls the data and the processing. Prominent examples of centralized systems are deep learning systems [8] and database management systems [11].

While finding wide adoption, the centralized paradigm has shortcomings. First and foremost, it is evident that clients of a centralized data system need to *trust* the entity controlling the system. Furthermore, by allowing the centralized system to control and process their data, clients give up their *autonomy* and *privacy*. Finally, the centralized entity can become a *single point of failure* if it stops operating the data system for whatever reason (e.g., technical reasons or economical reasons.) To overcome these shortcomings, it is common that researchers develop *decentralized* alternatives to centralized data systems. This phenomenon is not new, but has already been observed decades ago [9].

More recently, two promising instantiations of the decentralized data systems paradigm have received a lot of attention: Blockchain systems for transaction processing and federated learning systems for the training of deep neural networks. Blockchains like Bitcoin [10], Ethereum [17], or Hyperledger Fabric [1] allow for the decentralized processing and validation of transactions by a group of peer nodes. The transactions are organized as a sequence of immutable blocks that link each other by a cryptographic hash function. Federated learning systems like Flower [2] and TensorFlow Federated [3] allow clients to jointly train a deep neural network while keeping their data private [13]. This is achieved by only sharing model parameters that are aggregated after each training round instead of transferring the raw training data to a central entity. While tackling different use cases, both types of system have in common that they overcome centralization by putting control (over transaction processing or neural network training, respectively) into the hands of a large group of peer entities instead of centralizing it at a single party.

2 CHALLENGES

Despite their advantages, decentralized data systems face a couple of challenges:

Performance: First and foremost, performance of decentralized systems is typically lower than that of their centralized counterparts. This is due to the overhead induced by the need for coordination between the many peer nodes. As a consequence, blockchains suffer of a higher transaction latency than centralized databases, which can result in lower throughput and a higher rate of transaction conflicts [4]. Federated learning systems show a lower convergence speed and cause a higher energy consumption than centralized implementations [15].

Security. The central role of peer nodes in a decentralized data system opens new attack vectors. *Sybil attacks* try to influence the global decisions by infiltrating a large number of peer nodes into the system that are in fact all controlled by a single entity [6]. *Man-in-the-middle attacks* can attack single peer nodes by intercepting their communication with the rest of the system [7]. *Poisoning attacks* aim at introducing malicious data into the decentralized data system [12, 14]. There are numerous counter-measures against these attacks, but they are notoriously expensive [15].

Governance. Decentralized systems are hard to control, which makes governance difficult. In blockchain systems, this regards the changing of central system policies, such as the block size or the endorsement policy in Hyperledger Fabric [4, 5]. In federated learning, the quality of a trained neural network depends on the quality

Proceedings of the VLDB Endowment. ISSN 2150-8097.

of the training data. Hence, data governance plays an important role. Indeed, regulations like the EU AI Act put strict requirements on data governance into place [15]. However, the decentralized nature of federated learning makes data governance difficult, as all training data stays private on the peer nodes and is not accessible by a central entity.

Responsibility. In relevant legislation such as the GDPR and the EU AI Act, there is a clear assignment of responsibilities to a central entity that is in control of the data and processing. However, in a decentralized data system, it is unclear who is responsible and accountable for what part of the process. This makes it difficult to map the legal responsibilities to concrete, individual entities [16]. As no single entity is in control of the entire data system, there is the risk that either a single entity becomes accountable for actions that were out of its control, or that the entire system operates in a legal "gray area."

3 RESEARCH OPPORTUNITIES

In my keynote, I argue that the data management community is in a good position to tackle the challenges of decentralized data systems. However, current research is biased toward optimizing performance and security. While these questions are important, there is only little research performed on governance and responsibility.

In blockchain systems, legislation-related issues, such as GDPR's right to erasure and right to rectification vis-à-vis an immutable ledger, remain to be solved. Furthermore, there are open philosophical and ethical questions, such as the legitimation of a decentralized, anonymous group of peers to control business-critical transactions, or the implementation of ethics such as fairness and non-discrimination into blockchain systems. For federated learning systems, the most pressing question is that of data governance in a decentralized system [15]. Further issues regard human oversight, robustness and the need for auditing the system.

To tackle these kind of problems, an interdisciplinary research effort will be necessary. It will be important that the data management community engages in this discussion, as we have the technical expertise that is urgently needed.

REFERENCES

[1] Elli Androulaki, Artem Barger, Vita Bortnikov, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Christopher Ferris, Gennady Laventman, Yacov Manevich, Srinivasan Muralidharan, Chet Murthy, Binh Nguyen, Manish Sethi, Gari Singh, Keith Smith, Alessandro Sorniotti, Chrysoula Stathakopoulou, Marko Vukolić, Sharon Weed Cocco, and Jason Yellick. 2018. Hyperledger fabric: a distributed operating system for permissioned blockchains. In *Proceedings of the Thirteenth EuroSys Conference* (Porto, Portugal) *(EuroSys '18)*. Association for Computing Machinery, New York, NY, USA, Article 30, 15 pages. https://doi.org/10.1145/3190508.3190538

[2] Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. 2022. Flower: A Friendly Federated Learning Research Framework. arXiv:2007.14390 [cs.LG] https://arxiv.org/abs/2007.14390

[3] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. 2019. Towards Federated Learning at Scale: System Design. In *Proceedings of Machine Learning and Systems*, A. Talwalkar, V. Smith, and M. Zaharia (Eds.), Vol. 1. 374–388. https://proceedings.mlsys.org/paper_files/paper/2019/file/7b770da633baf74895be22a8807f1a8f-Paper.pdf

[4] Jeeta Ann Chacko, Ruben Mayer, and Hans-Arno Jacobsen. 2021. Why Do My Blockchain Transactions Fail?: A Study of Hyperledger Fabric. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava (Eds.). ACM, 221–234. https://doi.org/10.1145/3448016.3452823

[5] Jeeta Ann Chacko, Ruben Mayer, and Hans-Arno Jacobsen. 2023. How To Optimize My Blockchain? A Multi-Level Recommendation Approach. *Proc. ACM Manag. Data* 1, 1, Article 24 (may 2023), 27 pages. https://doi.org/10.1145/3588704

[6] John R Douceur. 2002. The sybil attack. In *International workshop on peer-to-peer systems*. Springer, 251–260.

[7] Parinya Ekparinya, Vincent Gramoli, and Guillaume Jourjon. 2018. Impact of Man-In-The-Middle Attacks on Ethereum. In *2018 IEEE 37th Symposium on Reliable Distributed Systems (SRDS)*. 11–20. https://doi.org/10.1109/SRDS.2018.00012

[8] Ruben Mayer and Hans-Arno Jacobsen. 2020. Scalable Deep Learning on Distributed Infrastructures: Challenges, Techniques, and Tools. *ACM Comput. Surv.* 53, 1, Article 3 (feb 2020), 37 pages. https://doi.org/10.1145/3363554

[9] Dejan S Milojicic, Vana Kalogeraki, Rajan Lukose, Kiran Nagaraja, Jim Pruyne, Bruno Richard, Sami Rollins, and Zhichen Xu. 2002. *Peer-to-Peer Computing*. Technical Report HPL-2002-57, HP Labs.

[10] Satoshi Nakamoto. 2008. Bitcoin: A peer-to-peer electronic cash system. http://bitcoin.org/bitcoin.pdf

[11] Raghu Ramakrishnan and Johannes Gehrke. 2002. *Database Management Systems*. McGraw-Hill, Inc.

[12] Teppei Sato, Mitsuyoshi Imamura, and Kazumasa Omote. 2020. Threat Analysis of Poisoning Attack Against Ethereum Blockchain. In *Information Security Theory and Practice*, Maryline Laurent and Thanassis Giannetsos (Eds.). Springer International Publishing, Cham, 139–154.

[13] René Schwermer, Ruben Mayer, and Hans-Arno Jacobsen. 2024. Federated Computing – Survey on Building Blocks, Extensions and Systems. arXiv:2404.02779 [cs.LG] https://arxiv.org/abs/2404.02779

[14] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. 2022. A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *ACM Comput. Surv.* 55, 8, Article 166 (dec 2022), 35 pages. https://doi.org/10.1145/3551636

[15] Herbert Woisetschläger, Alexander Erben, Bill Marino, Shiqiang Wang, Nicholas D. Lane, Ruben Mayer, and Hans-Arno Jacobsen. 2024. Federated Learning Priorities Under the European Union Artificial Intelligence Act. *CoRR* abs/2402.05968 (2024). https://doi.org/10.48550/ARXIV.2402.05968 arXiv:2402.05968

[16] Herbert Woisetschläger, Simon Mertel, Christoph Krönke, Ruben Mayer, and Hans-Arno Jacobsen. 2024. Federated Learning and AI Regulation in the European Union: Who is Responsible? – An Interdisciplinary Analysis. arXiv:2407.08105 [cs.AI] https://arxiv.org/abs/2407.08105

[17] Gavin Wood et al. 2014. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper* 151, 2014 (2014), 1–32.