New Approaches to Managing Metadata at Scale in Research Libraries

Timothy A. Thompson
Indiana University Bloomington
School of Informatics, Computing, and Engineering
Bloomington, Indiana 47408
timathom@indiana.edu

ABSTRACT

The analysis of big data often relies on distributed storage and computation; however, access to big data—and to the platforms capable of managing and processing it—continues to be largely centralized. Centralization is particularly evident in the case of the metadata produced, managed, and disseminated by academic and research libraries. Libraries typically create and share their catalog records by uploading them to a centrally managed database, which can then be searched by other libraries for records that can be copied and added to an institution's local catalog. This centralized approach, which operates on the basis of membership fees, has the advantage of scalability and availability, but it comes at the cost of a loss of autonomy. Although technical innovation is possible within the current paradigm, the growing maturity of peer-to-peer protocols and decentralized solutions points toward an alternative approach, one that would allow libraries to share their data directly without having to pay an expensive intermediary.

KEYWORDS

i523, hid-sp18-705, Research Libraries, Library Catalogs, Peer-to-Peer, Blockchain

1 INTRODUCTION

The problem of entity resolution (also known as record linkage or data matching [5]) is one that has a direct impact on the work of information professionals in research libraries. In library units responsible for catalog management, many workflows center on a procedure known as copy cataloging, which aims to expedite the processing of new acquisitions. Copy cataloging involves searching a shared database for records created by another cataloging agency, but that describe identical publications that have been acquired by one's local institution [6]. In the current environment, a single company, the Online Computer Library Center (OCLC—http://www.oclc.org), is the only viable platform for global cooperative cataloging [19]. OCLC provides data aggregation and warehousing services that allow libraries to effectively share their data, but its business model does not encourage peer-to-peer interaction and innovation among individual libraries. This vendor-driven paradigm entails the acceptance of a business model that, in effect, charges libraries for serving their own data back to them, with some added value through quality control and normalization. Once a library's data has been sent to OCLC, it also becomes subject to potential licensing restrictions, as well as the expectation that future dissemination of the data will include attribution of OCLC [12, 14].

2 NEW APPROACHES TO METADATA MANAGEMENT

Libraries have a tradition of experience with record matching and automation [11], but now stand to benefit from the increasingly mainstream availability of algorithms and routines developed within the context of data science and machine learning. Sophisticated algorithms for string comparison and probabilistic record linkage have long been available, but are not widely used by libraries, with the exception of large-scale projects such as the Social Networks and Archival Context Project (SNAC) (http://snaccooperative.org/) and the Virtual International Authority File (VIAF) (http://viaf.org/). The former has employed methods based on Naive Bayes classification algorithms to aggregate and disambiguate data from across a wide range of libraries and archives (the reported accuracy of the approach fell with the range of 80-90 percent) [7]. More recent approaches to record matching have improved on probabilistic methods such as Naive Bayes by using Artificial Neural Networks, improving accuracy rates in some cases to 98 percent or more [16].

As machine learning tools and methods have become more accessible, however, large-scale, real-time access to library metadata has not necessarily followed suit. The catalog of a large academic library may contain around 10 million records [20]. By comparison, as of August 2018, the OCLC catalog database, WorldCat, contained 427,501,671 bibliographic records in 491 languages [13]. As long as service providers such as OCLC maintain centralized control over the aggregated metadata of research libraries, large-scale computational analysis—and the innovation it could produce—will remain proprietary and locked away.

Although decentralization may be appealing as an ideal, librarians who manage bibliographic metadata are immersed in a discourse that centers on the idea of control: they use terms such as authority control, controlled vocabularies, and intellectual and physical control of collections [15]. The idea of control is closely related to the idea of trust: when workflows and systems are centralized, it becomes easier to enforce norms and standards, but it also becomes more likely that potential contributors may be excluded, especially when they are unable to afford the price of membership in a proprietary system.

New distributed technologies and protocols, including blockchains and distributed hash tables (DHTs), could allow research libraries to form robust peer-to-peer networks that would enable data sharing on a larger scale. Although public blockchains such as Ethereum and Bitcoin are limited in the amount of data that can feasibly be stored on chain, alternative platforms that address this limitation have recently emerged. The blockchain-based database service

BigchainDB, implemented in Python, provides a robust storage data solution while preserving the benefits of blockchain features such as data immutability and an asset-based transactional model. By running a consortial blockchain network of BigchainDB nodes, libraries could be empowered to abandon centralized models and begin managing their data collectively.

3 BLOCKCHAINS FOR RESEARCH LIBRARIES

Some in the library profession have been skeptical of blockchain applications for their domain, arguing that they have been overhyped as a panacea, when in reality they are nothing more than slow, expensive, append-only databases [1]. Even core developers working to support the Bitcoin blockchain have argued that the constraints imposed by blockchain technology, such as immutability and decentralized consensus, make it appropriate for a very limited set of applications—namely, currency and the exchange of value [18]. For individuals and organizations who are investigating blockchains as a technical solution, it is important from the outset to establish a framework for evaluating their applicability and appropriateness [17]. For example, a blockchain-based solution may be appropriate in a scenario in which there is a lack of trust among participants, or in which processes and collaboration would be more efficient if the need for trust were eliminated [17]. In the case of a shared catalog for research libraries, trust is an issue because not all participants can be trusted to provide data that conforms to expected levels of quality. A commercial, centralized solution mitigates these concerns by requiring participants to pay a membership fee. A blockchain solution addresses issues of trust by enforcing a decentralized consensus mechanism, which may take different forms, but which is designed to ensure that participants can trust the network to maintain a consistent state across all transactions [3].

The Proof-of-Stake consensus algorithm, employed by some blockchain networks as an alternative to Bitcoin's resource-intensive Proof-of-Work mechanism, is similar to the membership fee model in that validator nodes are elected based on their share of "stake" in the network, measured by their allocation of network tokens [8]. For research library applications, a variation of Proof-of-Stake known as Proof-of-Authority may be most appropiate solution [4, 8]. In contrast to public blockchains such as Ethereum and Bitcoin, or fully private blockchains restricted to a single organization, so-called consortium blockchains may be the preferred approach, one in which consensus "is controlled by a pre-selected set of nodes" [4]. The model implemented by the BigchainDB project fits the parameters of a consortium blockchain that implements a Proof-of-Authority approach to consensus [9].

4 DESIGN

A blockchain-based library catalog should support the creation of a decentralized marketplace for library metadata. Rather than paying a centralized exchange to distribute their catalog records, libraries could buy and sell records in a peer-to-peer exchange. Catalog records could thus become a source of revenue rather than a costly expenditure. Many blockchain systems support the creation of "smart assets," or the creation of tokens to represent real-word assets.

5 ARCHITECTURE

5.1 BigchainDB

5.1.1 The Evolution of BigchainDB.

5.1.2 BigchainDB Server. As of version 2.0, BigchainDB is Byzantine Fault Tolerant.

In BigchainDB 2.0, as is the case in general with systems that are Byzantine fault tolerant, $3f + 1$ nodes are necessary to run a network, where f is the number of faulty nodes to be tolerated [10].

5.1.3 Tendermint.

5.1.4 MongoDB.

6 DATASET

7 IMPLEMENTATION

8 CONCLUSION

Shown in Figure 1. Shown in Figure 2. Shown in Figure 3.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

9 CONCLUSION

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the i523 teaching assistants for their support and suggestions in writing this report.

REFERENCES

[1] Sue Alman. 2018. Blockchain: Apps and Ideas. Blockchains for the Information Profession Website. (Jul 2018). https://ischoolblogs.sjsu.edu/blockchains/blockchain-apps-and-ideas/ Accessed 2018.

[2] A.M. Antonopoulos. 2017. *Mastering Bitcoin: Programming the Open Blockchain* (2 ed.). O'Reilly, Sebastopol, CA, United States.

[3] Ethan Buchman, Jae Kwon, and Zarko Milosevic. 2018. The Latest Gossip on BFT Consensus. (2018). arXiv:cs.DC/1807.04938 https://arxiv.org/abs/1807.04938v2

[4] Vitalik Buterin. 2015. On Public and Private Blockchains. Ethereum Blog. (Aug 2015). https://blog.ethereum.org/2015/08/07/on-public-and-private-blockchains/ Accessed 2018.

[5] Peter Christen. 2012. *Data Matching*. Springer, Berlin, Germany.

[6] Claire Doran and Cheryl Martin. 2017. Measuring Success in Outsourced Cataloging: A Data-Driven Investigation. *Cataloging & Classification Quarterly* 55, 5 (2017), 307–317. https://doi.org/10.1080/01639374.2017.1317309

[7] R. R. Larson and K. Janakiraman. 2011. Connecting Archival Collections: The Social Networks and Archival Context Project. In *Research and Advanced Technology for Digital Libraries, TPDL 2011*. Springer, Berlin, 3–14. https://doi.org/10.1007/978-3-642-24469-8-3

[8] Gautier Marin. 2018. Understanding the Value Proposition of Cosmos. Cosmos Medium Blog. (Apr 2018). https://blog.cosmos.network/understanding-the-value-proposition-of-cosmos-ecaef63350d Accessed 2018.

[9] Troy McConaghy. 2018. [Reply in bigchaindb/bigchaindb Gitter Chat]. BigchainDB Gitter Chat. (Jun 2018). https://gitter.im/bigchaindb/bigchaindb?at=5b16ac9599fa7f4c0648cc13 Accessed 2018.

[10] Troy McConaghy. 2018. [Reply in bigchaindb/bigchaindb Gitter Chat]. BigchainDB Gitter Chat. (May 2018). https://gitter.im/bigchaindb/bigchaindb?at=5b055eaf9ed336150ea41180 Accessed 2018.

[11] Judy McQueen. 1992. Record Matching: Computers Cannot See That Which Is Obvious to "Any Idiot" . . . and Vice Versa. *Information Today* 9, 11 (Dec 1992), 41–44.

[12] OCLC. 2010. WorldCat Rights and Responsibilities. (Jun 2010). https://www.oclc.org/en/worldcat/cooperative-quality/policy.html Accessed 2018.

[13] OCLC. 2018. Inside WorldCat. (2018). https://www.oclc.org/en/worldcat/inside-worldcat.html Accessed 2018.

[14] OCLC. no date. WorldCat Data Licensing. (no date). https://www.oclc.org/content/dam/oclc/worldcat/documents/worldcat-data-licensing.pdf Accessed 2018.

[15] Hope A. Olson. 2001. The Power to Name: Representation in Library Catalogs. *Signs* 26, 3 (2001), 639–668. http://www.jstor.org/stable/3175535

[16] Orion F. Reyes-Galaviz, Witold Pedrycz, Ziyue He, and Nick J. Pizzi. 2017. A Supervised Gradient-Based Learning Algorithm for Optimized Entity Resolution. *Data & Knowledge Engineering* 112 (2017), 106–129. https://doi.org/10.1016/j.datak.2017.10.004

[17] B. A. Scriber. 2018. A Framework for Determining Blockchain Applicability. *IEEE Software* 35, 4 (Jul/Aug 2018), 70–77. https://doi.org/10.1109/MS.2018.2801552

[18] Jimmy Song. 2018. Why Blockchain Is Hard. Cryptocurrency Medium Blog. (May 2018). https://medium.com/@jimmysong/why-blockchain-is-hard-60416ea4c5c Accessed 2018.

[19] Amy H. Turner. 2010. OCLC WorldCat as a Cooperative Catalog. *Cataloging & Classification Quarterly* 48, 2-3 (2010), 271–278. https://doi.org/10.1080/01639370903536237 arXiv:https://doi.org/10.1080/01639370903536237

[20] Yale University Library. 2018. YUL Quicksearch Search Results. Yale University Library Quicksearch RSS Feed. (Oct 2018). http://search.library.yale.edu/catalog?commit=Search&format=atom&q=&search_field=all_fields

A CHKTEX

```
Already up to date.
WARNING: line longer than 80 characters
('    ', '105:', '%Taking this a step farther and also thinking about it in the context of linked data, which breaks down \n')
WARNING: line longer than 80 characters
('    ', '106:', '%metadata to the level of individual statements, or triples of subject-predicate-object, one could begin \n')
WARNING: line longer than 80 characters
('    ', '102:', '%to think about the valuation of metadata in different ways. For example, how much is a triple for a \n')
WARNING: line longer than 80 characters
('    ', '99:', '%subject heading worth in relation to a triple for a publication statement? The former presumably \n')
WARNING: line longer than 80 characters
('    ', '81:', '%requires more expertise to create, whereas the later can simply be transcribed.\n')
Warning 38 in content.tex line 176: You should not use punctuation in front of quotes.
systems support the creation of ``smart assets,'' or the creation of tokens
                                              ^
```

List of Figures

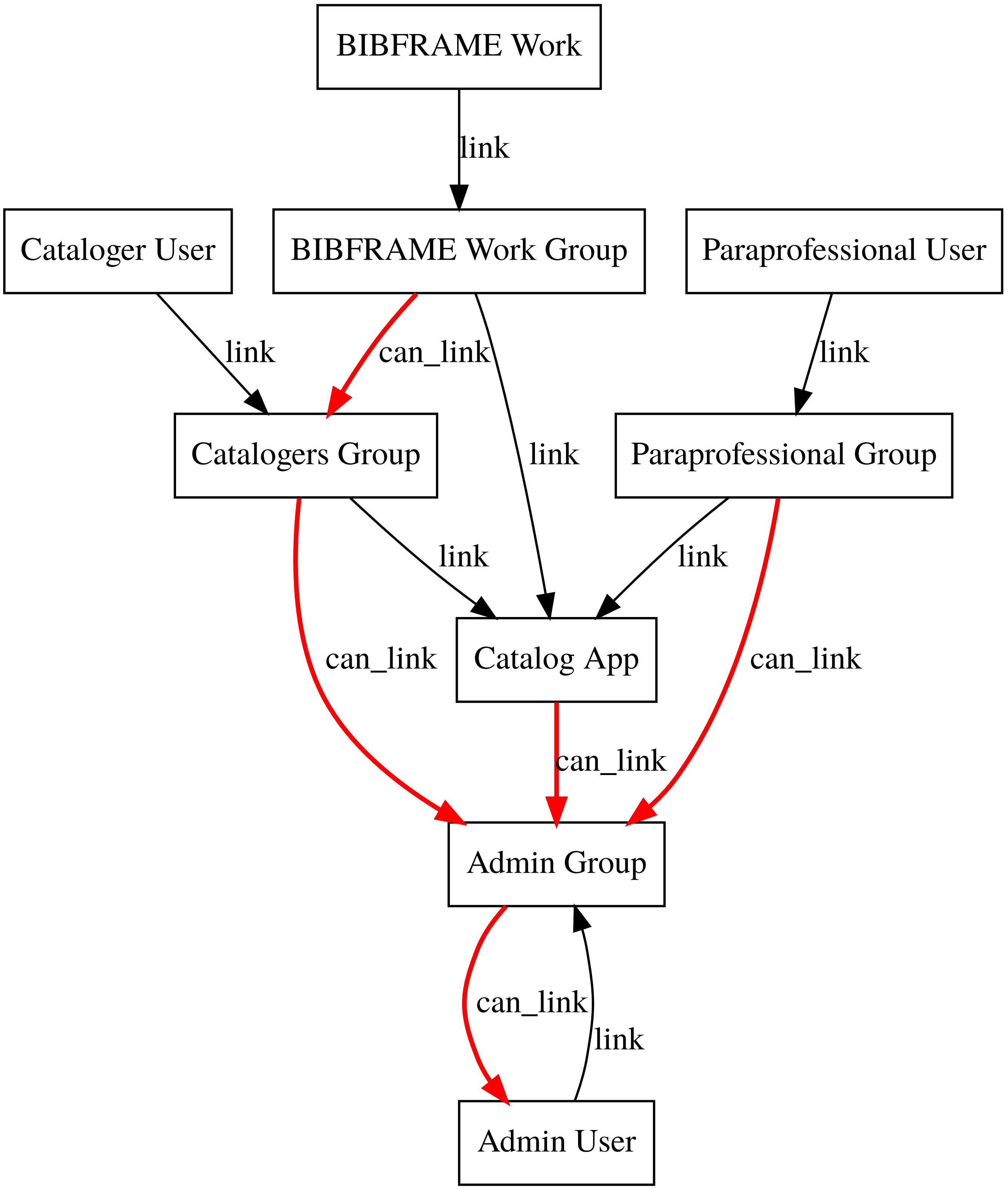

Figure 1: Graph of permissions in BigchainDB using Role-Based Access Control

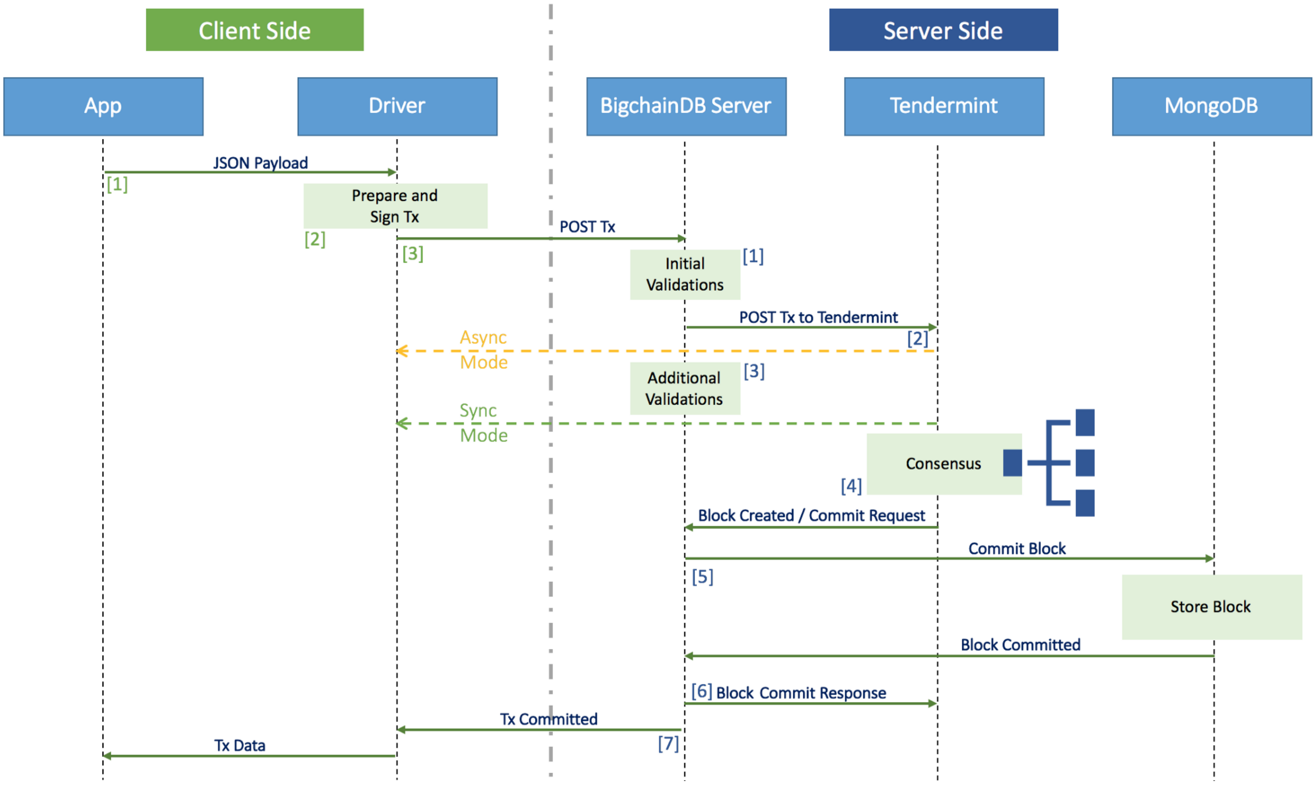

Figure 2: BigchainDB Sequence Diagram [2]

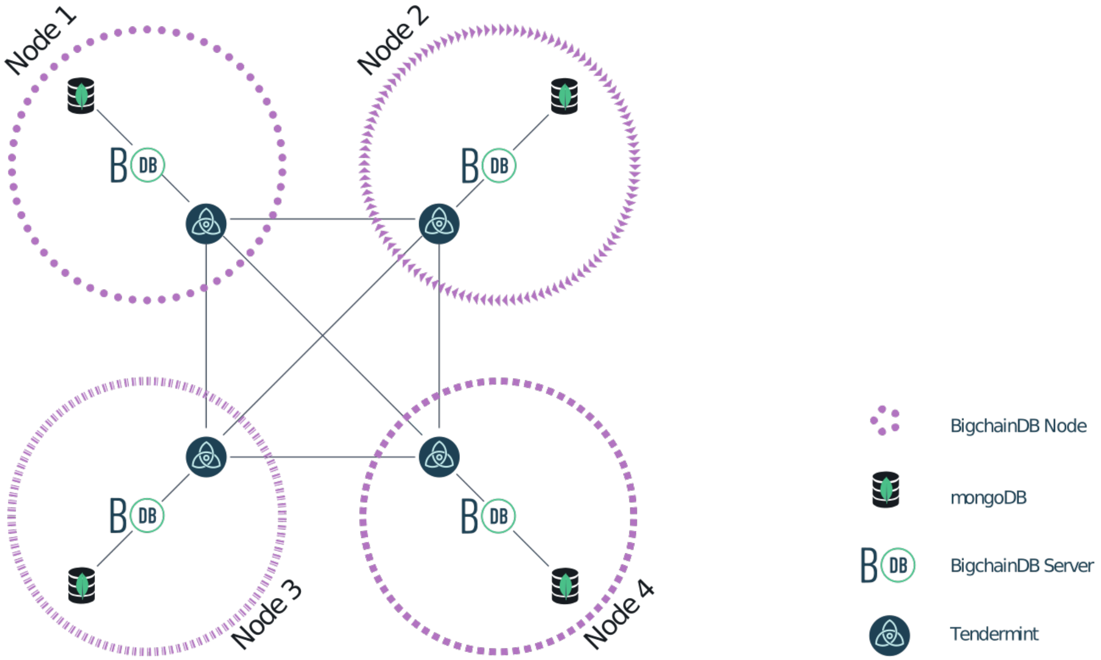

Figure 3: BigchainDB Architecture Diagram [2]