Telco Big Data Research and Open Problems

Constantinos Costa
Dept. of Computer Science
University of Pittsburgh
Pittsburgh, PA 15213, USA
costa.c@cs.pitt.edu

Demetrios Zeinalipour-Yazti
Dept. of Computer Science
University of Cyprus
1678 Nicosia, Cyprus
dzeina@cs.ucy.ac.cy

***Abstract*—A telecommunication company (telco) is traditionally only perceived as the entity that provides telecommunication services, such as telephony and data communication access to users. However, the radio and backbone infrastructure of such entities spanning densely most urban spaces and widely most rural areas, provides nowadays a unique opportunity to collect immense amounts of data that capture a variety of natural phenomena on an ongoing basis, e.g., traffic, commerce, mobility patterns and user service experience. The ability to perform analytics on the generated big data within tolerable elapsed time and share it with key smart city enablers (e.g., municipalities, public services, startups, authorities, and companies), elevates the role of telcos in the realm of future smart cities from pure network access providers to information providers. In this tutorial, we overview the state-of-the-art in telco big data analytics by focusing on a set of basic pillars, namely: (i) background and respective architectures; (ii) real-time analytics and detection; (iii) experience, behavior and retention analytics; (iv) privacy; and (v) storage. We also present experiences from developing an innovative such architecture and conclude with open problems and future directions.**

Keywords*—*Telco, Big Data, Queries, Analytics, Storage, Privacy

I. Introduction

A *telecommunication company (telco)* is the entity that provides telecommunication services, such as telephony and data communication access to users. The rapid expansion of broadband mobile networks, the pervasiveness of smartphones, and the introduction of dedicated Narrow Band connections for smart devices and Internet of Things (NB-IoT) [1], [2] have contributed to the expansion of the radio and backbone infrastructure of such entities in a way that these nowadays span densely most urban spaces and widely most rural areas. This has lead to the generation of very large amounts and variety of spatiotemporal *big* data that encapsulate a wide range of natural phenomena on an ongoing basis, e.g., traffic, commerce, mobility patterns and user service experience.

Consider a telco in the city of Shenzhen, China, which serves 10 million users and produces 5TB per day [3] (i.e., thousands to millions of records every second). Huang et al. [4] break their 2.26TB per day *Telco Big Data (TBD)* down as follows: (i) *Business Supporting Systems (BSS)* data, which is generated by the internal work-flows of a telco (e.g., billing, support), accounting to a moderate of 24GB per day and; (ii) *Operation Supporting Systems (OSS)* data, which is generated by the Radio and Core equipment of a telco, accounting to 2.2TB per day and occupying over 97% of the total volume.

Data exploration queries over such TBD are of great interest to both the telco operators and the smart city enablers (e.g., municipalities, public services, startups, authorities, and companies), as these allow for interactive analysis at various granularities, narrowing it down for a variety of tasks. Effectively storing and processing TBD workflows can unlock a wide spectrum of challenges, ranging from churn prediction of subscribers [4], city localization [5], 5G network optimization / user-experience assessment [6]–[8], optimizing public transportation [9] and road traffic mapping [10], [11]. Data exploration and visualization might be the most important tools in the big data era [7], [9], [12]–[16], where decision support makers, ranging from CEOs to front-line support engineers, aim to draw valuable insights and conclusions visually.

Our tutorial will tackle the topic of telco big data from a wide range of perspectives: fundamentals, definitions, current state, academic & industrial perspective, reality & visionary scenarios as well as future challenges. The tutorial captures the big picture, such that interested researchers and practitioners can expand their study by following the references. Our presentation is carried out through the lens of an experimental TBD architecture we developed at the University of Cyprus, coined SPATE [7], which is a SPAtio-TEmporal framework that uses both lossless data *compression* and lossy data *decaying* (i.e., *Data Postdiction* [17]) to ingest large quantities of TBD in the most compact manner.

Compression refers to the encoding of data using fewer bits than the original representation and is important as it shifts the resource bottlenecks from storage- and network-I/O to CPU, whose cycles are increasing at a much faster pace [18]–[20]. It also enables data exploration tasks to retain full resolution over the most important collected data. *Decaying* on the other hand, as suggested in [21], refers to the progressive loss of detail in information as data ages with time until it has completely disappeared (the schema of the database does not decay [22]). This enables data exploration tasks to retain high-level data exploration capabilities for predefined aggregates over long time windows, without consuming enormous amounts of storage.

Our tutorial aims to provide an extensive coverage of TBD research, which falls under the following categories: (i) background on TBD and respective architectures; (ii) real-time analytics and detection; (iii) experience, behavior and retention analytics; (iv) privacy; and (v) storage. There is also traditional telco research not related to big data, rather comprises of topics related to business (BSS) data in relational databases. The given presentation should allow the audience to grasp basic and advanced concepts ranging from the anatomy of a telco network and the structure of TBD all the way up to applications and benefits of TBD. We will conclude the tutorial with the presentation of the challenges and opportunities in the field.

To our knowledge, this is the first tutorial covering explicitly telco big data and this stems directly from our recent work on the subject covered in [7] [23] [17] and the TBD Awareness project[1]. The tutorial builts upon the experiences and preliminary feedback we got from an earlier presentation of the tutorial at IEEE MDM'18 [24]. We believe that the tutorial will be informative to both academics and practitioners creating vibrant discussions at the conference, as it deals with an application domain that will receive further attention in the future as societies move closer to the target of 5G networks.

OUTLINE

In this section we outline the tentative structure of the tutorial during the conference. The final layout of the tutorial will be reflected in its presentation available through the tutorial website [2].

What is Telco Big Data? In this first introductory section we will discuss a reference TBD architecture and focus on its respective data sources and basic storage and processing pillars. Particularly, we will start out by overviewing various telco big data sources, ranging from CDR to Mobile BroadBand (MBB), i.e., Network Measurement System (NMS) data or the CHR (Call Trace History) traffic Measurement Recording feature in Huawei's RNC and NodeB. We then overview telco big data identifiers (ICCID, IMSI, IMEI, MSISDN) and a respective relational diagram that links together the individual data sources. We will have also have a closer look on the Key Performance Indicators (KPIs) for a particular type of TBD analytics (i.e., churn prediction), to get a better sense of the multiple dimensions envolved in the acquisition of telco streams. Our discussion in the first section will conclude with an overview of reference architectures from both the academia and the industry.

Real-time Analytics and Detection: Zhang et al. [3] developed *OceanRT*, which was one of the first real-time TBD analytic demonstrations. Yuan et al. [25] present *OceanST* which features: (i) an efficient loading mechanism of ever-growing telco MBB data; (ii) new spatiotemporal index structures to process exact and approximate spatiotemporal aggregate queries. Iyer et al. [6] present *CellIQ* to optimize queries such as *"spatiotemporal traffic hotspots"* and *"hand-off sequences with performance problems"*. It represents the snapshots of cellular network data as graphs and leverages on the spatial and temporal locality of cellular network data. Zhu et al. [5] deal with the usage of telco MR data for city-scale localization, which is complementary to the scope of our work.

Braun et al. [26] develop a scalable distributed system that efficiently processes mixed workloads to answer event stream and analytic queries over telco data. Bouillet et al. [27] develop a system on top of IBM's InfoSphere Streams middleware that analyzes 6 billion CDR per day in real-time. Abbasoğlu et al. [28] present a system for maintaining call profiles of customers in a streaming setting by applying scalable distributed stream processing.

Experience, Behavior and Retention Analytics: Huang et al. [4] empirically demonstrate that customer churn prediction performance can be significantly improved with TBD. Although BSS data have been utilized in churn prediction very well in the past decade, the authors show how with a primitive Random Forest classifier TBD can improve churn prediction accuracy from 68% to 95%. Luo et al. [8] propose a framework to predict user behavior involving more than one million telco users. They represent users as documents containing a collection of changing spatiotemporal "words" that express user behavior. By extracting the users' space-time access records from MBB data, they learn user-specific compact topic features that they use for user activity level prediction. Ho et. al. [29] propose a distributed community detection algorithm that aims to discover groups of users that share similar edge properties reflecting customer behavior.

Privacy: Hu et al. [30] study Differential Privacy for data mining applications over TBD and show that for real-word industrial data mining systems the strong privacy guarantees given by differential privacy are traded with a 15% to 30% loss of accuracy. Privacy and confidentiality are critical for telcos' reliability due to the highly sensitive attributes of user data located in CDR, such as billing records, calling numbers, call duration, data sessions, and trajectory information. *SPATE* deals with privacy-aware data sharing as a functionality for next generation smart-city applications.

Storage: Telcos are reaching a point where they are collecting more data than they could possibly exploit. This has the following two implications: (i) it introduces a significant financial burden on the operator to store the collected data locally. Notice that the deep storage of data in public clouds, where economies-of-scale are available (e.g., AWS Glacier), is not an option due to privacy reasons; and (ii) it imposes a high computational cost for accessing and processing the collected data. Consequently, we claim that the vision of infinitely storing all IoT-generated velocity data on fast high-availability or even deep storage will gradually become too costly and impractical for many analytic-oriented processing scenarios.

From a telco's perspective, the requirement is to: (i) *incrementally store big data in the most compact manner*, and (ii) *improve the response time for data exploration queries*. These two objectives are naturally conflicting, as conjectured in [31]. In previous work, custom data management systems have been designed with the objectives to save storage space using compression, and speed up temporal range queries using indices [32]–[35]. None of these considers the notion of "decay" as expressed in [21], which suggests sacrificing either accuracy or read efficiency for less frequently accessed data to save space. We will explore these directions as part of the SPATE architecture [7] we developed and the *TBD-DP (Data Postdiction)* operator [17]. Unlike data prediction, which aims to make a statement about the future value of some tuple in a TBD store, data postdiction aims to make a statement about the past value of some tuple that does not exist anymore, as it had to be deleted to free up space. TBD-DP relies on existing Machine Learning (ML) algorithms to abstract TBD into compact models that can be stored and queried when necessary. Our proposed TBD-DP operator has the following two conceptual phases: (i) in an offline phase, it utilizes a LSTM-based hierarchical ML algorithm to learn a tree of models (coined TBD-DP tree) over time and space; (ii) in

[1]TBD Awareness. https://tbd.cs.ucy.ac.cy/

[2]Tutorial slides: https://dmsl.cs.ucy.ac.cy/tutorials/icde19/

an online phase, it uses the TBD-DP tree to recover data with a certain accuracy. We claim that the LSTM model is capturing the essence of the past through its short and long-term dependencies, similarly to how the brain retains both recent information and important old information at a high resolution.

II. Audience, Relevance, History and Duration

The goal of this tutorial is to convey a basic and advanced understanding of the unique characteristics, challenges and opportunities of telco big data management and how these can facilitate data engineering research and applications.

Audience: The tutorial is targeted to scientists with a basic understanding of data management, but no knowledge of indoor data management or telco big data management technologies is required. In particular, this tutorial addresses the following audience:

- Graduate and Undergraduate Students
- Data Engineering Researchers/Educators
- Industry Developers

Relevance: This tutorial covers, but is not limited to, the following ICDE 2019 topics of interest:

- Data Integration, Metadata Management, and Interoperability
- Data Stream Systems and Sensor Networks
- Data Visualization and Interactive Data Exploration
- Database Privacy, Security, and Trust
- Distributed, Parallel and P2P Data Management
- Database technology for machine learning
- Query Processing, Indexing, and Optimization
- Temporal, Spatial, Mobile and Multimedia

History: An earlier version of our tutorial has appeared at the 19th IEEE Intl. Conference on Mobile Data Management on June 28, 2018 in Aalborg, Denmark (https://dmsl.cs.ucy.ac.cy/tutorials/mdm18/), where it was well attended and attracted a lot of interest from participants. The basic aim of repeating this extended tutorial at IEEE ICDE'19 is to publicize the material of telco big data management to a wider target audience that might capitalize upon the findings and initiate new research and development projects.

Duration: 1 hour and 30 minutes.

Biographies of Tutorial Presenters

Constantinos Costa is a Visiting Lecturer at the Department of Computer Science at University of Pittsburgh, PA, USA and a Research Associate at the Advanced Data Management Technologies Laboratory (ADMT). His primary research interests include Spatial Big Data Management, particularly distributed query processing for spatial and spatio-temporal datasets. He holds a Ph.D. in Computer Science (2018) from the University of Cyprus and his thesis was titled "Algorithms and Indexing Structures for Spatial Big Data". Besides research, he has distinguished in several programming and innovation competitions and is an active contributor to several industrial and open-source systems for telco big data, indoor navigation, crowd messaging. He has industrial experience in the telco big data sector. For more information please visit: https://www.cs.ucy.ac.cy/~costa.c/.

Demetrios Zeinalipour-Yazti is an Associate Professor of Computer Science at the University of Cyprus, where he leads the Data Management Systems Laboratory (DMSL). His primary research interests include Data Management in Computer Systems and Networks, particularly Mobile and Sensor Data Management; Big Data Management in Parallel and Distributed Architectures; Spatio-Temporal Data Management; Network and Telco Data Management; Crowd, Web 2.0 and Indoor Data Management; Data Privacy Management. He holds a Ph.D. in Computer Science from University of California - Riverside (2005). Before his current appointment, he served the University of Cyprus as an Assistant Professor and Lecturer but also the Open University of Cyprus as a Lecturer. He has held visiting research appointments at Akamai Technologies, Cambridge, MA, USA, the University of Athens, Greece, the University of Pittsburgh, PA, USA and the Max Planck Institute for Informatics, Saarbrücken, Germany. He is a Humboldt Fellow, Marie-Curie Fellow, an ACM Distinguished Speaker (2017-2020), a Senior Member of ACM, a Senior Member of IEEE and a Member of USENIX. He serves on the editorial board of Distr. and Par. Databases (Elsevier), Big Data Research (Springer) and is an independent evaluator for the European Commission (Marie Skodowska-Curie and COST actions).

His h-index is 24, holds over 2800 citations, has an Erdös number of 3, won 10 international awards (ACMD17, ACMS16, IEEES16, HUMBOLDT16, IPSN14, EVARILOS14, APPCAMPUS13, MDM12, MC07, CIC06) and delivered over 30 invited talks. He has participated in over 20 projects funded by the US National Science Foundation, by the European Commission, the Cyprus Research Promotion Foundation, the Univ. of Cyprus, the Open University of Cyprus and the Alexander von Humboldt Foundation, Germany. Finally, he has also been involved in industrial Research and Development projects (e.g., Finland, Taiwan and Cyprus) and has technically lead several mobile data management services (e.g., Anyplace, Rayzit and Smartlab) reaching over 35K users worldwide with over 140K sessions. For more information please visit: https://www.cs.ucy.ac.cy/~dzeina/ or the DMSL website: https://dmsl.cs.ucy.ac.cy/.

References

[1] Ericsson.com, "Cellular Networks For Massive IoT enabling low power wide area applications," 2016. [Online]. Available: https://goo.gl/Sf2Cj4

[2] C. Byrne, "Fast data use cases for telecommunications," in *O'Reilly Media*. Inc. Release Date: October 2017, ISBN: 9781491998267, 2017. [Online]. Available: https://www.oreilly.com/library/view/fast-data-use/9781491998267/

[3] S. Zhang, Y. Yang, W. Fan, L. Lan, and M. Yuan, "Oceanrt: Real-time analytics over large temporal data," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '14. New York, NY, USA: ACM, 2014, pp. 1099–1102.

[4] Y. Huang, F. Zhu, M. Yuan, K. Deng, Y. Li, B. Ni, W. Dai, Q. Yang, and J. Zeng, "Telco churn prediction with big data," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD. New York, NY, USA: ACM, 2015, pp. 607–618.

[5] F. Zhu, C. Luo, M. Yuan, Y. Zhu, Z. Zhang, T. Gu, K. Deng, W. Rao, and J. Zeng, "City-scale localization with telco big data," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ser. CIKM. New York, NY, USA: ACM, 2016, pp. 439–448.

[6] A. P. Iyer, L. E. Li, and I. Stoica, "Celliq: Real-time cellular network analytics at scale," in *Proceedings of the 12th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'15. Berkeley, CA, USA: USENIX Association, 2015, pp. 309–322.

[7] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Efficient exploration of telco big data with compression and decaying," in *33rd IEEE International Conference on Data Engineering, ICDE 2017, San Diego, CA, USA, April 19-22, 2017*, 2017, pp. 1332–1343. [Online]. Available: https://doi.org/10.1109/ICDE.2017.175

[8] C. Luo, J. Zeng, M. Yuan, W. Dai, and Q. Yang, "Telco user activity level prediction with massive mobile broadband data," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, pp. 63:1–63:30, May 2016.

[9] G. Di Lorenzo, M. Sbodio, F. Calabrese, M. Berlingerio, F. Pinelli, and R. Nair, "Allaboard: Visual exploration of cellphone mobility data to optimise public transport," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 2, pp. 1036–1050, Feb 2016.

[10] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Towards real-time road traffic analytics using telco big data," in *Proceedings of the International Workshop on Real-Time Business Intelligence and Analytics, BIRTE, Munich, Germany, August 28, 2017*, 2017, pp. 5:1–5:5. [Online]. Available: http://doi.acm.org/10.1145/3129292.3129296

[11] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson, "Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones," in *Proceedings of the 7th International Conference on Embedded Networked Sensor Systems, SenSys 2009, Berkeley, California, USA, November 4-6, 2009*, 2009, pp. 85–98. [Online]. Available: https://doi.org/10.1145/1644038.1644048

[12] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact." *MIS quarterly*, vol. 36, no. 4, pp. 1165–1188, 2012.

[13] TeraLab, "TeraLab Data Science for Europe," 2016. [Online]. Available: http://www.teralab-datascience.fr/

[14] A. Eldawy, M. F. Mokbel, S. Alharthi, A. Alzaidy, K. Tarek, and S. Ghani, "Shahed: A mapreduce-based system for querying and visualizing spatio-temporal satellite data," in *2015 IEEE 31st International Conference on Data Engineering*, April 2015, pp. 1585–1596.

[15] Y. Zheng, W. Wu, H. Zeng, N. Cao, H. Qu, M. Yuan, J. Zeng, and L. M. Ni, "Telcoflow: Visual exploration of collective behaviors based on telco data," in *2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016*, 2016, pp. 843–852. [Online]. Available: https://doi.org/10.1109/BigData.2016.7840677

[16] W. Wu, J. Xu, H. Zeng, Y. Zheng, H. Qu, B. Ni, M. Yuan, and L. M. Ni, "Telcovis: Visual exploration of co-occurrence in urban human mobility based on telco data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 935–944, Jan 2016.

[17] C. Costa, A. Charalampous, A. Konstantinidis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Decaying telco big data with data postdiction," in *19th IEEE International Conference on Mobile Data Management, MDM 2018, Aalborg, Denmark, June 25-28, 2018*, 2018, pp. 106–115. [Online]. Available: https://doi.org/10.1109/MDM.2018.00027

[18] Y. Chen, A. Ganapathi, and R. H. Katz, "To compress or not to compress - compute vs. io tradeoffs for mapreduce energy efficiency," in *Proceedings of the First ACM SIGCOMM Workshop on Green Networking*, ser. Green Networking '10, 2010, pp. 23–28.

[19] B. Welton, D. Kimpe, J. Cope, C. M. Patrick, K. Iskra, and R. Ross, "Improving i/o forwarding throughput with data compression," in *2011 IEEE Intl. Conference on Cluster Computing*, Sept 2011, pp. 438–445.

[20] T. Bicer, J. Yin, D. Chiu, G. Agrawal, and K. Schuchardt, "Integrating online compression to accelerate large-scale data analytics applications," in *Parallel Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on*, May 2013, pp. 1205–1216.

[21] M. L. Kersten, "Big data space fungus," in *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*, 2015.

[22] M. Stonebraker, R. Castro, F. Dong Deng, and M. Brodie, "Database decay and what to do about it." 2016. [Online]. Available: https://goo.gl/tJNa9m

[23] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Towards real-time road traffic analytics using telco big data," in *Proceedings of the 11th Intl. Workshop on Real-Time Business Intelligence and Analytics, collocated with VLDB 2017*, ser. BIRTE'17. August 28, 2017, Munich, Germany: ACM International Conference Proceedings Series, 2017, conference, pp. 5:1–5:5. [Online]. Available: http://db.cs.pitt.edu/birte2017/

[24] C. Costa and D. Zeinalipour-Yazti, "Telco big data: Current state & future directions," in *19th IEEE International Conference on Mobile Data Management, MDM 2018, Aalborg, Denmark, June 25-28, 2018*, 2018, pp. 11–14. [Online]. Available: https://doi.org/10.1109/MDM.2018.00016

[25] M. Yuan, K. Deng, J. Zeng, Y. Li, B. Ni, X. He, F. Wang, W. Dai, and Q. Yang, "Oceanst: A distributed analytic system for large-scale spatiotemporal mobile broadband data," *Proc. VLDB Endow.*, vol. 7, no. 13, pp. 1561–1564, Aug. 2014.

[26] L. Braun, T. Etter, G. Gasparis, M. Kaufmann, D. Kossmann, D. Widmer, A. Avitzur, A. Iliopoulos, E. Levy, and N. Liang, "Analytics in motion: High performance event-processing and real-time analytics in the same database," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD. New York, NY, USA: ACM, 2015, pp. 251–264.

[27] E. Bouillet, R. Kothari, V. Kumar, L. Mignet, S. Nathan, A. Ranganathan, D. S. Turaga, O. Udrea, and O. Verscheure, "Processing 6 billion cdrs/day: From research to production (experience report)," in *Proceedings of the 6th ACM Intl. Conference on Distributed Event-Based Systems*, ser. DEBS, 2012, pp. 264–267.

[28] M. A. Abbasoğlu, B. Gedik, and H. Ferhatosmanoğlu, "Aggregate profile clustering for telco analytics," *Proc. VLDB Endow.*, vol. 6, no. 12, pp. 1234–1237, Aug. 2013.

[29] Q. Ho, W. Lin, E. Shaham, S. Krishnaswamy, T. A. Dang, J. Wang, I. C. Zhongyan, and A. She-Nash, "A distributed graph algorithm for discovering unique behavioral groups from large-scale telco data," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ser. CIKM. New York, NY, USA: ACM, 2016, pp. 1353–1362.

[30] X. Hu, M. Yuan, J. Yao, Y. Deng, L. Chen, Q. Yang, H. Guan, and J. Zeng, "Differential privacy in telco big data platform," *Proc. VLDB Endow.*, vol. 8, no. 12, pp. 1692–1703, Aug. 2015.

[31] M. Athanassoulis, M. S. Kester, L. M. Maas, R. Stoica, S. Idreos, A. Ailamaki, and M. Callaghan, "Designing access methods: The rum conjecture," in *Intl. Conf. on Ext. Database Technology (EDBT)*, 2016.

[32] S. Lakshminarasimhan, N. Shah, S. Ethier, S. Klasky, R. Latham, R. Ross, and N. F. Samatova, "Compressing the incompressible with isabela: In-situ reduction of spatio-temporal data," in *European Conference on Parallel Processing*. Springer, 2011, pp. 366–379.

[33] E. R. Schendel, Y. Jin, N. Shah, J. Chen, C.-S. Chang, S.-H. Ku, S. Ethier, S. Klasky, R. Latham, R. Ross *et al.*, "Isobar preconditioner for effective and high-throughput lossless data compression," in *IEEE 28th Intl. Conference on Data Engineering*, 2012, pp. 138–149.

[34] J. Jenkins, I. Arkatkar, S. Lakshminarasimhan, D. A. Boyuka II, E. R. Schendel, N. Shah, S. Ethier, C.-S. Chang, J. Chen, H. Kolla *et al.*, "Alacrity: Analytics-driven lossless data compression for rapid in-situ indexing, storing, and querying," in *Transactions on Large-Scale Data- and Knowledge-Centered Systems X*. Springer, 2013, pp. 95–114.

[35] E. Soroush and M. Balazinska, "Time travel in a scientific array database," in *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*. IEEE, 2013, pp. 98–109.